2005-05-06

A Bayesian Approach to Detect the Onset of Activity Limitation Among Adults in NHIS

Yan Bai
Worcester Polytechnic Institute

A Bayesian Approach to Detect the Onset of Activity Limitation

Among Adults in NHIS

by

Yan Bai

A Thesis

Submitted to the Faculty

of

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Master of Science

in

Aplied Statistics

May 2005

APPROVED:

Dr. Balgobin Nandram, Thesis Advisor

Dr. Bogdan Vernescu, Department Head

Abstract

Data from the 1995 National Health Interview Survey (NHIS) indicate that, due to chronic conditions, the onset of activity limitation typically occurs between age 40-70 years (i.e., the proportion of young adults with activity limitation is small and roughly constant with age and then it starts to change, roughly increasing). We use a Bayesian hierarchical model to detect the change point of a positive activity limitation status (ALS) across twelve domains based on race, gender, and education.

We have two types of data: weighted and unweighted. We obtain weighted binomial counts using a regression analysis with the sample weights. Given the proportion of individuals in the population with positive ALS, we assume that the number of individuals with positive ALS at each age group has a binomial probability mass function. The proportions across age are different, and have the same beta distribution up to the change point (unknown), and the proportions after the change point have a different beta distribution.

We consider two different analyses. The first considers each domain individually in its own model and the second considers the twelve domains simultaneously in a single model to "borrow strength" as in small area estimation. It is reasonable to assume that each domain has its own onset.In the first analysis, we use the Gibbs sampler to fit the model, and a computation of the marginal likelihoods, using an output analysis from the Gibbs sampler, provides the posterior distribution of the change point. We note that a reversible jump sampler fails in this analysis because it tends to get stuck either age 40 or age 70. In the second analysis, we use the Gibbs sampler to fit only the joint posterior distribution of the twelve change points. This is a difficult problem because the joint density requires the numerical computation of a triple integral at each iteration. The other parameters of the process are obtained using data augmentation by a Metropolis sampler and a Rao-Blackwellization.

We found that overall the age of onset is about 50 to 60 years.

Acknowledgements

I extend my sincere gratitude and appreciation to many people who made this masters thesis possible. First of all, I am highly indebted to my advisor, Dr. Balgobin Nandram, for his great help, guidiance, patience, and understanding. It is an exciting and rewarding experience to work with such a knowledgeable and kind professor to catch up with his endless ideas and creativity.

Thanks are due to Dr. Jai W. Choi of the National Center for Health Statistics (NCHS) for his encouragement and assistance throughout the whole project. I am also grateful to Mr. Douglas Williams, who helped me to get an internship at NCHS during summer 2004.

A special thank you to my parents and dear husband, for their forever love and support.

Chapter 1

Introduction

The National Health Interview Survey (NHIS) is an important source of imformation on the health of the U.S. population. One of the variables in NHIS is activity limitation status (ALS). Activity limitation among adults due to chronic conditions is a major health problem in United States. It may influence the quality of an individual's life, and it can cause socio-economic problems. We believe that the activity limitation status changes at certain age, probably between 40 and 70 (see Figure 1). The main issue we want to address in this study is to find the onset of activity limitation among adults.

1.1 National Health Interview Survey

The National Health Interview Survey (NHIS) has been conducted every year since 1957 by the National Center for Health Statistics (NCHS) to measure an aspect of health status of the U.S. population [1]. Through this sample survey, NCHS conducts surveys on chronic and acute conditions, doctor visits, hospital episodes, disability, household and personal information, and other special aspects of health of the U.S. population.

The questionnaire is divided up into two major sections, core and supplemental. The core section includes items on household and personal information, basic health questions on conditions, doctor visit, hospital discharge and other supplemental information. The supplemental section includes questions about selected interests from the general public, encompassing a wide range of topics such as prescription medicine, hypertension, diabetes, high blood pressure, and

HIV. The core section is administered annually and the supplemental section is administered as its need arises.

1.2 Activity Limitation Status

Respondents in NHIS were asked to provide their activity limitation status during the interview. The degree of activity limitation is divided in four categories: 1. Unable to perform major activity. 2. Limited in kind/amount of major activity. 3. Limited in other activities. 4. Not limited (includes unknowns). Here, major activity refers to activities like going to work, going to school, keeping house, etc.

Loss of activities of daily living will lead to disability. After the onset of disability, an individual's life may be influenced both physically and psychologically [2]. Because of the disability, they may encounter changes in paid and unpaid work, and therefore, have lower income and less benefit. On the other hand, they will need more help in their daily living and the cost of living will increase. These will result in higher risk of poverty. While they are having all these economical difficulties, they may also have social problems. Disability may prevent people from normal social life and make them socially excluded. Disability is assessed using a longitudinal study. The main issue in this study is activity limitation, which is different from disability. However, knowing what impact of disability would make the onset of activity limitation meaningful. Doctors can give patients suggestions when they are near the onset of activity limitation, therefore patients can prevent the activity limitation or be prepared for the impact of it.

1.3 Description of Data

The data we used for this study is 1995 National Health Interview Survey. The interviewed sample was has 41,824 households containing 102,467 persons. The range of age is from 0 to 99. Since we are only interested in the onset of activity limitation among adults with chronic conditions, we only use the data from those whose age is from 30 to 80. The variable Activity Limitation Status (ALS) has 4 levels, and we recoded it into 2 levels in 2 ways. In the first case,

we group the first two levels together and give it value 1, which indicate that the individual have some major activity limitation; and level 3 and 4 are also grouped together and given the value 0, which means the individual have almost no activity limitation. In the second case, we group level 2, 3, 4 together, and give it value 0, which indicate that the individual has no severe major activity limitation; and we leave level 1 as it is and give it value 1, which indicates that the individual has major activity limitation.

Previous studies indicate that characteristics such as race, gender, and socioeconomic status influence the probability of having functional disorders [10]. Socioeconomic status include factors such as education and income. Education is determined early in life and influence psychosocial mechanisms throughout life. It is highly correlated with income. However, education is more strongly predictive of onset of functional health problem such as activity limitation, while income is more predictive of course or progression. Activity limitation in our study is a chronic health problem of adulthood, and is the outcome of a long process of development as a function of exposure to a wide range of social, psychological, and biomedical risk factors. Education indexes both the socio-economic position of individuals early in adulthood and a stock of human capital available to them from that time on. All these influence long-term patterns of exposure to and experience of major psychosocial and biomedical risk factors that cause activity limitation. Income, on the other hand, is usually measured in year, and thus reflects socioeconomic position and resources closer to the time of assessment of activity limitation. It is more strongly related to the resources available for the treatment or management of the health problem. Therefore, income more relates to the severity of the problem while education more strongly relates to the onset or existence of the health problem. So in our study, we also include variables like race, gender, and education. There are two levels of race, white and nonwhite. Gender also have two levels, male and female. In the original data, there are 7 levels of education. To fit our model, it is recoded in to 3 levels, 1. pre-college, 2. college, and 3. post-college. Each combination of race, gender, and education is considered as a domain, therefore, the dataset is divided into 12 domains and one can expect different onset for each domain. All data combined is considered as the 13^{th} domain. The boxplots of the proportions of positive ALS for the 12 domains by age are presented in Figure 1 & 2. One can notice that

there is a sudden drop of proportions of ALS around age 70. We believe that this is because many people with positive ALS are of poor health, and they pass away around age 70, therefore, the proportion of positive ALS drops.

Since the NHIS uses a multistage sample designed to sample the population of the United States, it is necessary to utilise the person's basic weight for proper analysis of the data. The weight for each sample person is the product of four component weights:

1. **Probability of selection.** The basic weight for each person is obtained by multiplying the reciprocals of the probabilities of selection at each step in the design: PSU, segment, and household.

2. **Household nonresponse adjustment within segment.** In the NHIS, interviews are completed in about 94 percent of all eligible households. Because of household nonresponse, a weighting adjustment is required. The nonresponse adjustment weight is a ratio of the within-segment weighted number of sample households divided by the within- segment weighted number of actually interviewed households, both numbers exclusive of households with unknown black/Hispanic status. For segments with nonresponding households of unknown black/Hispanic status, the previously mentioned factor was multiplied by the ratio of the number of segment households divided by the number of known status households. This adjustment reduces bias in an estimate to the extent that persons in the noninterviewed households have the same characteristics as the persons in the interviewed households in the same segment.

3. **First-stage ratio adjustment.** The weight for persons in the nonself-representing PSU's is ratio adjusted to the 1990 population within four race-residence classes of the nonself-representing strata within each geographic region.

4. **Poststratification by age-sex-race-ethnicity.** Within each of 88 age-sex-race-ethnicity cells, a weight is constructed each quarter to ratio adjust the first-stage population estimate based on the NHIS to an independent estimate of the population of each cell. These independent estimates are prepared by the U.S. Bureau of the Census and are updated quarterly.

In addition to the design and ratio adjustments included in the person basic weight, the person weight is further modified depending on the variable selected, the length of the recall period, and the period of time for which the estimate is to be made. For a review of weighting

methods, see Kalton and Florence-Cervantes [8]. To construct the weighted data, first let y_{ijl} denote the value of ALS for the l^{th} individual in i^{th} domain and j years old, and let w_{ijl} be the weight of ALS for the l^{th} individual in i^{th} domain and j years old, then we define $\hat{p}_{ij}$ as the following,

$$\hat{p}_{ij} = \frac{\sum_{l \in D_{ij}} w_{ijl} y_{ijl}}{\sum_{l \in D_{ij}} w_{ijl}} \quad i = 1, \ldots, 12, \quad j = 30, \ldots, 80, \tag{1}$$

where D_{ij} are domains formed by race, sex, and education. Then we fit a logistic regression based on these $\hat{p}_{ij}$,

$$\log\left[\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right] = \underset{\sim}{X}'_{ij}\underset{\sim}{\beta} + \varepsilon_{ij}. \tag{2}$$

Using a first-order Taylor expansion, we get the variance,

$$Var\left(\log\left[\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right]\right) \approx \frac{1-p_{ij}}{p_{ij}} \frac{1-p_{ij}+p_{ij}}{(1-p_{ij})^2} = \frac{Var(\hat{p}_{ij})}{[p_{ij}(1-p_{ij})]^2}, \tag{3}$$

where

$$Var(\hat{p}_{ij}) = \frac{\sum_{l \in D_{ij}} w_{ijl}^2 p_{ij}(1-p_{ij})}{\left(\sum_{l \in D_{ij}} w_{ijl}\right)^2} = \frac{p_{ij}(1-p_{ij}) \sum_{l \in D_{ij}} w_{ijl}^2}{\left(\sum_{l \in D_{ij}} w_{ijl}\right)^2}. \tag{4}$$

$\underset{\sim}{X}_{ij}$ is a vector that contains an intercept and the value of sex, race, and eduation for each domain and each age group.

Now substituting (4) into (3) we get,

$$Var\left(\log\left[\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right]\right) \approx \frac{\sum_{l \in D_{ij}} w_{ijl}^2}{\hat{p}_{ij}(1-\hat{p}_{ij})\left(\sum_{l \in D_{ij}} w_{ijl}\right)^2} = \frac{1}{W_{ij}}. \tag{5}$$

Then the least square estimate for $\underset{\sim}{\beta}$ becomes

$$\hat{\underset{\sim}{\beta}} = \left\{\sum_i \sum_j W_{ij} \underset{\sim}{X}_{ij} \underset{\sim}{X}'_{ij}\right\}^{-1} \sum_i \sum_j W_{ij} \log\left[\frac{\hat{p}_{ij}}{1-\hat{p}_{ij}}\right] \underset{\sim}{X}_{ij}. \tag{6}$$

In the case in which $\hat{p}_{ij} = 0$ or 1, we add $1/2n_{ij}$ to both $\hat{p}_{ij}$ and $1 - \hat{p}_{ij}$. This is similar to the adjustment in Cox [4] for the empirical logistic transform. Thus, we construct the new weighted

binomial data

$$y_{ij} = [\frac{n_{ij} e^{\underset{\sim}{X}'_{ij}\hat{\underset{\sim}{\beta}}}}{1 + e^{\underset{\sim}{X}'_{ij}\hat{\underset{\sim}{\beta}}}}], \tag{7}$$

where [] is the nearest integer, $y_{ij} = 0, 1, \ldots, n_{ij}$.

The number of individuals with ALS using the unweighted, simple weighted, and logistic weighted data for each race, gender, and education domain is presented in Tables 1 and 2, together with the proportions of positive ALS for each case. We observe that the simple weighted data is similar with the unweighted data, so in our analysis, we will use both the unweighted and logistic weighted data. The boxplots for the proportion of adults with positive ALS using the logistic weighted data are presented in Figures 3-4. We notice that the plots are smoother than those using the unweighted data and the sudden drop around age 70 is removed.

Domain	Total	Observed	Obs per	Simple wt	Sw per	Logistic wt	Lw per
1	1361	329	.242	369	.271	360	.265
2	263	96	.365	96	.365	82	.312
3	1503	366	.244	403	.268	398	.265
4	322	119	.370	120	.373	108	.335
5	5814	854	.147	856	.147	894	.154
6	1129	229	.203	227	.201	230	.204
7	7023	999	.142	998	.142	1117	.159
8	1538	304	.198	302	.196	316	.205
9	6012	463	.077	461	.077	522	.087
10	934	89	.095	87	.093	99	.106
11	5969	519	.087	526	.088	489	.082
12	1236	116	.094	115	.093	128	.104

Table 1.1: Observed ALS, Simple weight ALS, Logistic Weight ALS, and their proportions, where only "Limited in major activity" is considered as positive ALS.

To have some rough idea about where the change point is, we can use the Expectation Maximization algorithm [5] to compute the distribution for change point k. Consider the following simple approximate model for each domain,

$$\begin{aligned} \hat{p}_j &\overset{iid}{\sim} \text{Normal}(\mu_1, \tfrac{\hat{p}_j(1-\hat{p}_j)}{n_j}) \qquad j = 30, \ldots, k \\ \hat{p}_j &\overset{iid}{\sim} \text{Normal}(\mu_2, \tfrac{\hat{p}_j(1-\hat{p}_j)}{n_j}) \qquad j = k+1, \ldots, 80, \end{aligned} \tag{8}$$

where $\hat{p}_j$ is the proportion of people with positive ALS who are j years old. Before the change point, $\hat{p}_j$'s are independently and identically normally distributed with mean μ_1 and variance

Domain	Total	Observed	Obs per	Simple wt	Sw per	Logistic wt	Lw per
1	1361	242	.178	280	.206	257	.189
2	263	76	.289	77	.293	67	.255
3	1503	206	.137	218	.145	235	.156
4	322	67	.208	65	.202	72	.224
5	5814	494	.085	491	.084	559	.096
6	1129	162	.143	158	.140	167	.148
7	7023	450	.064	447	.064	577	.082
8	1538	163	.106	162	.105	195	.127
9	6012	203	.034	203	.034	277	.046
10	934	48	.051	51	.055	63	.067
11	5969	236	.040	240	.040	211	.035
12	1236	57	.046	58	.047	61	.049

Table 1.2: Observed ALS, Simple weight ALS, Logistic Weight ALS, and their proportions, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

$\hat{p}_j(1-\hat{p}_j)/n_j$, where n_j is total number of individuals who are j years old. After the change point, $\hat{p}_j$'s follow the same distributions as before, except that the mean is μ_2.

Now we can apply the EM algorithm to find the distribution of k,

$$E(k|\hat{\mu}_1,\hat{\mu}_2,\hat{p}) = \frac{\sum_{k=40}^{70} k \prod_{j=30}^{80} \left[\frac{n_j}{2\pi\hat{p}_j(1-\hat{p}_j)}\right]^{\frac{1}{2}} e^{-\frac{1}{2}\left\{\frac{\sum_{j=30}^{k} n_j(\hat{p}_j-\hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)}+\frac{\sum_{j=k+1}^{80} n_j(\hat{p}_j-\hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)}\right\}}}{\sum_{k=40}^{70} \prod_{j=30}^{80} \left[\frac{n_j}{2\pi\hat{p}_j(1-\hat{p}_j)}\right]^{\frac{1}{2}} e^{-\frac{1}{2}\left\{\frac{\sum_{j=30}^{k} n_j(\hat{p}_j-\hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)}+\frac{\sum_{j=k+1}^{80} n_j(\hat{p}_j-\hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)}\right\}}} \quad (9)$$

$$= \frac{\sum_{k=40}^{70} k e^{-\frac{1}{2}\left\{\frac{\sum_{j=30}^{k} n_j(\hat{p}_j-\hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)}+\frac{\sum_{j=k+1}^{80} n_j(\hat{p}_j-\hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)}\right\}}}{\sum_{k=40}^{70} e^{-\frac{1}{2}\left\{\frac{\sum_{j=30}^{k} n_j(\hat{p}_j-\hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)}+\frac{\sum_{j=k+1}^{80} n_j(\hat{p}_j-\hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)}\right\}}}. \quad (10)$$

Then the distribution of k will be

$$P(k|\hat{\mu}_1,\hat{\mu}_2,\hat{p}) = \frac{e^{-\frac{1}{2}\left\{\frac{\sum_{j=30}^{k} n_j(\hat{p}_j-\hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)}+\frac{\sum_{j=k+1}^{80} n_j(\hat{p}_j-\hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)}\right\}}}{\sum_{k=40}^{70} e^{-\frac{1}{2}\left\{\frac{\sum_{j=30}^{k} n_j(\hat{p}_j-\hat{\mu}_1)^2}{\hat{p}_j(1-\hat{p}_j)}+\frac{\sum_{j=k+1}^{80} n_j(\hat{p}_j-\hat{\mu}_2)^2}{\hat{p}_j(1-\hat{p}_j)}\right\}}}, \quad (11)$$

where

$$\hat{\mu}_1 = \frac{\sum_{j=30}^{k} n_j\hat{p}_j}{\sum_{j=30}^{k} n_j} \qquad \hat{\mu}_2 = \frac{\sum_{j=k+1}^{80} n_j\hat{p}_j}{\sum_{j=k+1}^{80} n_j} \quad (12)$$

The results are presented in Tables 1.2 to 1.6. These are only rough and approximate estimates for the distributions for k, and we note that the distributions are quite concentrated.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	**.237**	.000	.001	.035	.228	.012	.204	.000	.000	.000	.000	.000	.000
41	.150	.000	.005	.024	.095	.002	.046	.000	.000	.000	.000	.000	.000
42	.118	.000	.001	.024	**.292**	.174	.066	.000	.000	.000	**.664**	.000	.000
43	.148	.000	.016	.024	.144	**.363**	**.418**	.000	.000	.000	.074	.000	.000
44	.098	.000	**.763**	.024	.106	.259	.089	.000	.000	.000	.003	.000	.000
45	.092	.000	.146	.034	.073	.062	.131	.000	.003	.000	.240	.000	.009
46	.092	.000	.043	.034	.028	.040	.019	.000	.000	.000	.008	.000	.000
47	.046	.000	.018	.034	.015	.023	.008	.000	.158	.000	.009	.000	.006
48	.014	.000	.006	.019	.011	.044	.005	.012	**.565**	.000	.001	.000	**.957**
49	.003	.000	.002	.029	.005	.003	.010	**.607**	.140	.003	.000	.000	.028
50	.000	.000	.001	.029	.002	.003	.003	.318	.025	.017	.001	.000	.000
51	.000	.000	.000	.039	.001	.008	.000	.048	.080	.017	.000	.001	.000
52	.000	.001	.000	.048	.001	.003	.000	.012	.018	.017	.000	.000	.000
53	.000	.000	.000	.056	.000	.001	.000	.002	.004	**.119**	.000	.000	.000
54	.000	.001	.000	.051	.000	.001	.000	.001	.003	.117	.000	.000	.000
55	.000	.001	.000	.057	.000	.001	.000	.000	.002	.103	.000	.032	.000
56	.000	.001	.000	.073	.000	.001	.000	.000	.002	.103	.000	.023	.000
57	.000	.003	.000	**.083**	.000	.000	.000	.000	.000	.103	.000	**.156**	.000
58	.000	.001	.000	.046	.000	.000	.000	.000	.000	.060	.000	**.156**	.000
59	.000	.001	.000	.046	.000	.000	.000	.000	.000	.066	.000	**.156**	.000
60	.000	.001	.000	.054	.000	.000	.000	.000	.000	.037	.000	.141	.000
61	.000	.001	.000	.062	.000	.000	.000	.000	.000	.037	.000	.100	.000
62	.000	.001	.000	.005	.000	.000	.000	.000	.000	.037	.000	.100	.000
63	.000	.001	.000	.002	.000	.000	.000	.000	.000	.046	.000	.029	.000
64	.000	.020	.000	.004	.000	.000	.000	.000	.000	.035	.000	.029	.000
65	.000	.018	.000	.008	.000	.000	.000	.000	.000	.027	.000	.020	.000
66	.000	.045	.000	.005	.000	.000	.000	.000	.000	.011	.000	.016	.000
67	.000	.060	.000	.007	.000	.000	.000	.000	.000	.011	.000	.015	.000
68	.000	.110	.000	.014	.000	.000	.000	.000	.000	.011	.000	.015	.000
69	.000	**.366**	.000	.017	.000	.000	.000	.000	.000	.011	.000	.004	.000
70	.000	**.366**	.000	.017	.000	.000	.000	.000	.000	.011	.000	.004	.000

Table 1.3: Simple estimates for the change point using the EM algorithm, where only "Limited in major activity" is considered as positive ALS and the data is unweighted.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	**.239**	.028	.000	.042	.043	.025	**.995**	.000	.000	.000	.000	**.885**	.000
41	.142	.026	.000	.052	.014	.002	.004	.000	.000	.000	.000	.093	.000
42	.176	.026	.000	.063	.259	.211	.000	.000	.000	.000	.002	.017	.000
43	.098	.026	**.409**	.048	**.454**	**.349**	.001	.000	.000	.000	.000	.003	.000
44	.139	.031	.409	.048	.165	.295	.000	.000	.004	.000	.000	.000	.000
45	.150	.031	.160	.028	.047	.071	.000	.000	**.982**	.000	**.972**	.000	.000
46	.043	.031	.019	.028	.011	.030	.000	.003	.006	.001	.018	.001	.000
47	.010	.031	.003	.024	.006	.009	.000	.004	.008	.007	.006	.000	.100
48	.001	.031	.000	**.128**	.001	.005	.000	**.902**	.000	.007	.001	.000	**.483**
49	.000	.030	.001	.080	.000	.000	.000	.059	.000	.001	.000	.000	.416
50	.000	.030	.000	.061	.000	.000	.000	.028	.000	.002	.000	.000	.000
51	.000	.031	.000	.024	.000	.002	.000	.004	.000	.017	.000	.000	.000
52	.000	.032	.000	.025	.000	.000	.000	.000	.000	.017	.000	.000	.000
53	.000	.029	.000	.021	.000	.000	.000	.000	.000	**.672**	.000	.000	.000
54	.000	.031	.000	.034	.000	.000	.000	.000	.000	.117	.000	.000	.000
55	.000	.027	.000	.034	.000	.000	.000	.000	.000	.013	.000	.000	.000
56	.000	.030	.000	.019	.000	.000	.000	.000	.000	.027	.000	.000	.000
57	.000	.035	.000	.023	.000	.000	.000	.000	.000	.027	.000	.000	.000
58	.000	.029	.000	.035	.000	.000	.000	.000	.000	.016	.000	.000	.000
59	.000	.029	.000	.043	.000	.000	.000	.000	.000	.026	.000	.000	.000
60	.000	.029	.000	.038	.000	.000	.000	.000	.000	.009	.000	.000	.000
61	.000	.029	.000	.029	.000	.000	.000	.000	.000	.005	.000	.000	.000
62	.000	.030	.000	.026	.000	.000	.000	.000	.000	.006	.000	.000	.000
63	.000	.033	.000	.031	.000	.000	.000	.000	.000	.013	.000	.000	.000
64	.000	.039	.000	.008	.000	.000	.000	.000	.000	.006	.000	.000	.000
65	.000	.039	.000	.003	.000	.000	.000	.000	.000	.003	.000	.000	.000
66	.000	.040	.000	.001	.000	.000	.000	.000	.000	.001	.000	.000	.000
67	.000	.039	.000	.001	.000	.000	.000	.000	.000	.001	.000	.000	.000
68	.000	.040	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000
69	.000	**.043**	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000
70	.000	**.043**	.000	.000	.000	.000	.000	.000	.000	.001	.000	.000	.000

Table 1.4: Simple estimates for the change point using the EM algorithm, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS and the data is unweighted.

Age	**D1**	**D2**	**D3**	**D4**	**D5**	**D6**	**D7**	**D8**	**D9**	**D10**	**D11**	**D12**	**D13**
40	.000	.004	.000	.006	.000	.010	.000	.001	.006	.008	.000	.001	.000
41	.001	.005	.000	.010	.000	.012	.000	.002	.011	.009	.000	.003	.000
42	.001	.005	.000	.010	.000	.014	.000	.004	.021	.015	.001	.003	.000
43	.002	.005	.000	.010	.000	.022	.000	.007	.041	.025	.001	.004	.000
44	.005	.005	.000	.011	.000	.028	.000	.012	.064	.024	.003	.007	.000
45	.008	.005	.000	.027	.001	.027	.000	.016	.091	.022	.008	.010	.000
46	.012	.006	.000	.027	.004	.035	.000	.028	.096	.028	.011	.009	.000
47	.018	.006	.001	.027	.008	.039	.001	.037	.101	.025	.020	.010	.000
48	.021	.006	.002	.067	.025	.041	.002	.040	**.107**	.025	.032	.011	.000
49	.032	.007	.005	.067	.039	.044	.005	.045	.083	.024	.033	.014	.000
50	.060	.007	.007	.061	.083	.056	.013	.054	.066	.038	.039	.026	.000
51	**.104**	.013	.008	.055	.132	**.069**	.024	.058	.062	.045	.055	.027	.000
52	.098	.013	.016	.073	.133	.060	.049	.058	.050	**.071**	.058	.035	.000
53	.086	.015	.033	.073	**.156**	.058	.096	**.062**	.047	.061	.058	.035	.005
54	.080	.026	.054	.079	.117	.057	**.135**	.061	.034	.052	.066	.033	.084
55	.074	.044	.061	.079	.118	.047	.121	.059	.024	.043	.056	.052	.131
56	.092	.047	.077	**.084**	.075	.055	.123	.049	.021	.042	.066	.052	**.446**
57	.077	.062	.077	.060	.046	.050	.122	.046	.018	.041	.065	**.054**	.268
58	.068	.059	.094	.049	.025	.041	.115	.047	.015	.041	.066	.051	.035
59	.057	.067	**.140**	.044	.014	.040	.098	.056	.012	.039	**.072**	.051	.030
60	.043	**.088**	.137	.028	.012	.027	.052	.049	.008	.036	.057	.051	.000
61	.025	.073	.109	.020	.006	.028	.028	.054	.006	.033	.046	.051	.000
62	.014	.083	.064	.010	.003	.026	.009	.045	.005	.030	.035	.051	.000
63	.010	.088	.043	.006	.002	.018	.003	.031	.003	.030	.027	.051	.000
64	.007	.068	.031	.006	.001	.019	.001	.027	.002	.030	.029	.045	.000
65	.003	.046	.017	.004	.000	.017	.000	.019	.002	.030	.028	.045	.000
66	.001	.038	.012	.003	.000	.014	.000	.016	.001	.030	.023	.045	.000
67	.000	.032	.008	.003	.000	.014	.000	.009	.001	.030	.017	.045	.000
68	.000	.034	.004	.002	.000	.012	.000	.005	.001	.026	.012	.045	.000
69	.000	.024	.002	.001	.000	.010	.000	.003	.000	.026	.009	.040	.000
70	.000	.018	.001	.001	.000	.008	.000	.002	.000	.023	.006	.040	.000

Table 1.5: Simple estimates for the change point using the EM algorithm, where only "Limited in major activity" is considered as positive ALS and the data is weighted.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.000	.003	.000	.001	.000	.002	.000	.001	.000	.035	.000	.010	.000
41	.000	.005	.000	.001	.000	.004	.000	.005	.000	**.045**	.000	.010	.000
42	.000	.005	.000	.001	.000	.008	.000	.008	.000	.043	.000	.015	.000
43	.000	.005	.000	.001	.000	.012	.000	.013	.000	.042	.000	.021	.000
44	.000	.008	.000	.003	.000	.014	.000	.027	.000	.039	.000	.031	.000
45	.000	.008	.001	.005	.001	.021	.000	.042	.002	.038	.001	.039	.000
46	.001	.010	.002	.006	.002	.031	.000	.064	.006	.043	.003	.040	.000
47	.002	.015	.005	.006	.008	.049	.000	.068	.022	.042	.010	.039	.000
48	.004	.015	.008	.011	.025	.058	.001	.074	.044	.040	.024	.034	.000
49	.006	.018	.015	.011	.048	.062	.005	.081	.066	.036	.044	.036	.000
50	.009	.023	.025	.015	.107	**.079**	.023	.077	.110	.040	.082	.036	.000
51	.017	.022	.060	.020	**.201**	.072	.061	.075	**.162**	.037	.108	.035	.000
52	.025	.027	.073	.020	.196	.071	.106	**.085**	.156	.041	**.152**	.041	.000
53	.050	.041	.089	.020	.149	.055	.184	.082	.134	.034	.114	.038	.000
54	.078	.039	**.135**	.039	.113	.059	**.190**	.057	.098	.028	.097	**.051**	.002
55	.120	.037	.120	.039	.063	.053	.145	.050	.066	.023	.098	.047	.008
56	.103	.046	.126	.076	.045	.060	.150	.041	.047	.027	.076	.043	.166
57	.125	.079	.094	.098	.024	.073	.078	.033	.034	.031	.068	.042	**.794**
58	.122	**.093**	.075	.077	.010	.054	.037	.037	.023	.027	.041	.037	.024
59	**.154**	.071	.064	**.102**	.004	.038	.016	.031	.014	.029	.028	.034	.005
60	.085	.056	.037	.078	.002	.027	.004	.018	.007	.030	.023	.033	.000
61	.053	.054	.029	.101	.001	.025	.001	.014	.004	.032	.012	.031	.000
62	.023	.041	.019	.077	.000	.021	.000	.007	.002	.033	.009	.031	.000
63	.011	.051	.011	.048	.000	.014	.000	.004	.001	.028	.005	.029	.000
64	.008	.047	.007	.047	.000	.013	.000	.003	.000	.027	.003	.032	.000
65	.003	.036	.002	.036	.000	.008	.000	.002	.000	.025	.001	.032	.000
66	.001	.035	.001	.023	.000	.006	.000	.001	.000	.023	.001	.030	.000
67	.000	.033	.000	.018	.000	.005	.000	.000	.000	.022	.000	.030	.000
68	.000	.041	.000	.011	.000	.004	.000	.000	.000	.021	.000	.028	.000
69	.000	.021	.000	.005	.000	.003	.000	.000	.000	.021	.000	.023	.000
70	.000	.016	.000	.004	.000	.002	.000	.000	.000	.020	.000	.021	.000

Table 1.6: Simple estimates for the change point using the EM algorithm, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS and the data is weighted.

1.4 Literature Review

A sequence of random variables, $y = (y_1, \ldots, y_n)$ is said to have a single change-point at k if their distribution function is $F_{\theta 1}(y)$ for $i \leq k$ and $F_{\theta 2}(y)$ for $i > k$, where $F_{\theta 1}(y)$ and $F_{\theta 2}(y)$ are different and unknown distributions belonging to the same parametric family. The problem of

estimating the location of the change-point k has been extensively studied for the past several decades. Smith [9] proposed a Bayesian approach to the inferences about the change-point based on the posterior probabilities of the possible change-points. Assuming the density function is $p_1(x \mid \theta_1)$ and $p_2(x \mid \theta_2)$, the joint distribution of $x_1, \ldots, x_n$ conditional on θ_1, θ_2 and the change-point $k(1 \leq k \leq n)$ is given by

$$p(x_1, \ldots, x_n \mid k, \theta_1, \theta_2) = p_1(x_1, \ldots, x_k \mid \theta_1) p_2(x_{k-1}, \ldots, x_n \mid \theta_2) = \prod_{i=1}^{k} p_1(x_i \mid \theta_1) \prod_{i=k+1}^{n} p_2(x_i \mid \theta_2).$$

With θ_1 and θ_2 unknown, assuming a prior distribution of k over the set of possible change-point $p_0(k)(1 \leq k \leq n)$ such that $p_0(1)+p_0(2)+\cdots+p_0(n) = 1$ and a prior density of θ_1, θ_2, $p_0(\theta_1, \theta_2)$, Smith [9] obtained

$$p_n(k) \propto p(x_1, \ldots, x_n \mid k) p_0(k),$$

where

$$p(x_1, \ldots, x_n \mid k) = \int_{\Theta_{1,2}} p(x_1, \ldots, x_n \mid k, \theta_1, \theta_2) p_0(\theta_1, \theta_2) d\theta_1 d\theta_2.$$

Then inference about θ_1 and θ_2 can be made based on

$$p_n(\theta_1, \theta_2) = \sum_k p_n(\theta_1, \theta_2 \mid k) p_n(k),$$

where

$$p_n(\theta_1, \theta_2 \mid k) \propto p(x_1, \ldots, x_n \mid k, \theta_1, \theta_2) p_0(\theta_1, \theta_2).$$

Assuming uniform priors, the joint posterior that gives the posterior moments of $\hat{k}, \hat{\theta}_{1,\hat{k}}, \hat{\theta}_{2,\hat{k}}$, where $\hat{k}$ maximizes

$$p_1(x_1, \ldots, x_k \mid \hat{\theta}_{1,k}) p_2(x_{k-1}, \ldots, x_n \mid \hat{\theta}_{2,k}),$$

and $\hat{\theta}_{1,k}$ maximizes $p(x_1, \ldots, x_k \mid \theta_1)$. This general approach is applied to binomial and normal distributions and other situations such as θ_1 and θ_2 known are also discussed.

A later study [3] further explored the change-point problem using hierarchical Bayesian models. The desired marginal posterior densities are obtained utilizing the Gibbs sampler, a Markov chain Monte Carlo method. Suppose we have a collection of p random variables $U_1, \ldots, U_p$ whose full conditional probability denoted by $f(U_s|U_r, r \neq s), s = 1, \ldots, p$ are available for sampling. Under mild conditions, these full conditional distributions uniquely determine the full joint distribution $f(U_1, \ldots, U_p)$, and hence all marginal distributions $f(U_s), s = 1, \ldots, p$. The Gibbs sampler generates samples from the joint distribution as follows. Given an arbitrary starting set of values $U_1^{(0)}, \ldots, U_p^{(0)}$, we draw $U_1^{(1)}$ from $f(U_1|U_2^{(0)}, \ldots, U_p^{(0)})$, then $U_2^{(1)}$ from $f(U_2|U_1^{(0)}, U_3^{(0)}, \ldots, U_p^{(0)})$, and so on up to $U_p^{(1)}$ from $f(U_p|U_1^{(0)}, \ldots, U_{p-1}^{(0)})$. Under mild conditions, this p-tuple converges in distribution to a random observation from $f(U_1, \ldots, U_p)$ as $t \to \infty$. Replicate this process a large number of times and the samples then can be used for estimation of any of the marginal densities that we desire. The Gibbs sampler algorithm avoids sophisticated analytic and numerical high dimensional integration procedures, which makes some previously inaccessible problems doable. This approach can be applied to changing regressions, changing Poisson processes and changing Markov chains.

The study by [6] is an application of the approach. They used a Bayesian multinomial change-point analysis to determine the authorship of a book. A multinomial sequence of conditionally independent ordered random variables $y = (y_1, \ldots, y_n)$ is assumed for the number of words of different length in each chapter of the book. Different parameters, θ_a and θ_b are defined for the models before and after the change point respectively, each following a conjugate Dirichlet prior distribution. Then Bayesian hierachical models are fit and the posterior distribution of the change point can be obtained. Using the Gibbs sampler algorithm, inferences about the location of the change point and the multinomial parameters can be made. A Bayesian cluster analysis has also been applied and it confirms the existance of the changpoint. All these models assume that there is only one change-point, and there is only one parameter before the change-point and one parameter after the change-point.

However, the Markov chain Monte Carlo methods for Bayesian computation are restricted to problems where the joint distribution of all variables has a density with respect to some fixed standard underlying measure. They are not available for applications to Bayesian model

determination, where the dimensionality of the parameter vector is not fixed. To solve this problem, Green [7] proposed a new framework for the construction of reversible Markov chain samplers that jump between parameter subspaces of differing dimensionality, which is flexible and entirely constructive. When there are many competing models with different parameter spaces (e.g., dimensions), there is uncertainty about the model itself. A parameter can be created to index the models. All models are then fitted simultaneously, and the reversible jump sampler jumps over models.

1.5 A Brief Outline of the Theis

The rest of the thesis is as follows.

In Chapter 2 we show how to use a Bayesian model to estimate the change point for each domain seperately. This procedure is computationally intensive. So we consider an alternative procedure in Chapter 3, which uses the reversible jump sampler. Unfortunately, that procedure gets stuck at the boundaries of the parameter space, and therefore we cannot rely on it. For many of the domains, the change point is very different. Thus, we attempt to "borrow strength" across the domain. This procedure, in Chapter 4, is similar to the one in Chapter 2, but it is slightely more complex, because we need to integrate out all parameters to form a Gibbs sampler of the change points. Chapter 5 has some comparisons and diagnostics.

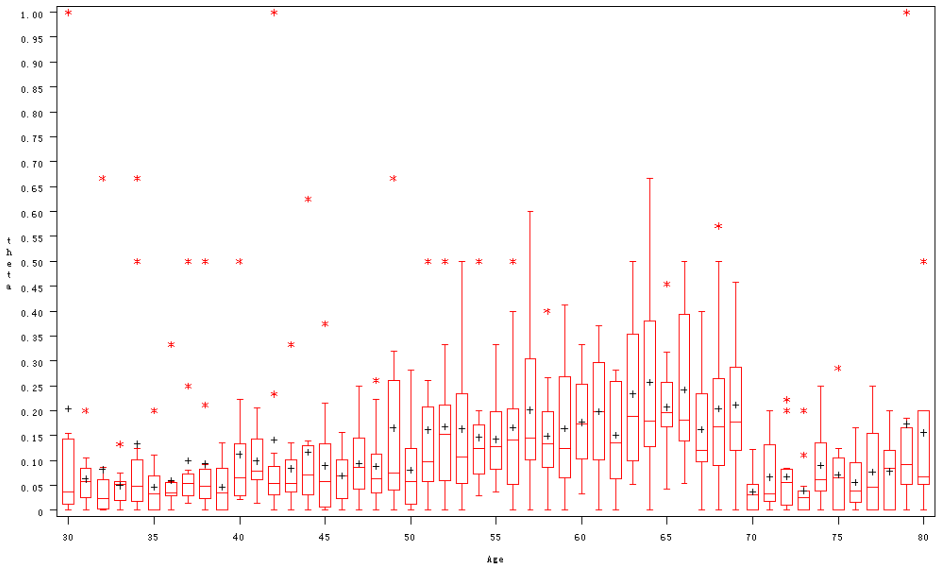

Figure 1.1: Boxplot of the proportions of positive ALS for the 12 domains by age using unweighted data, where only "Limited in major activity" is considered as positive ALS.

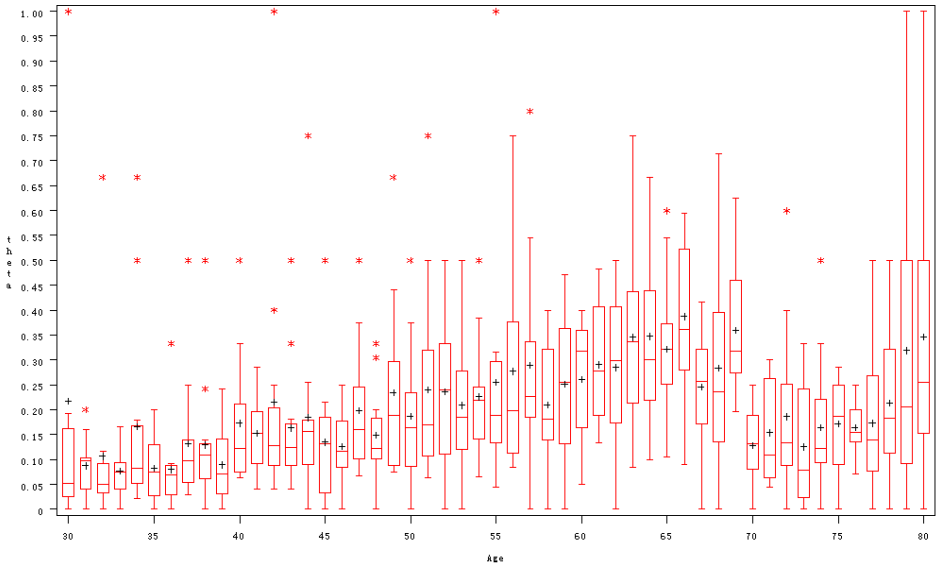

Figure 1.2: Boxplot of the proportions of positive ALS for the 12 domains by age using unweighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

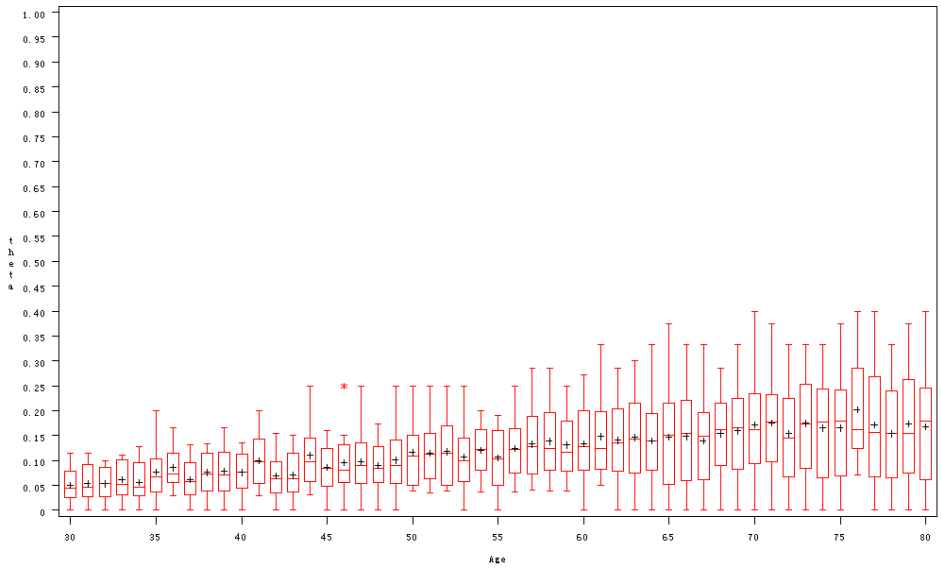

Figure 1.3: Boxplot of the proportions of positive ALS for the 12 domains by age using logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

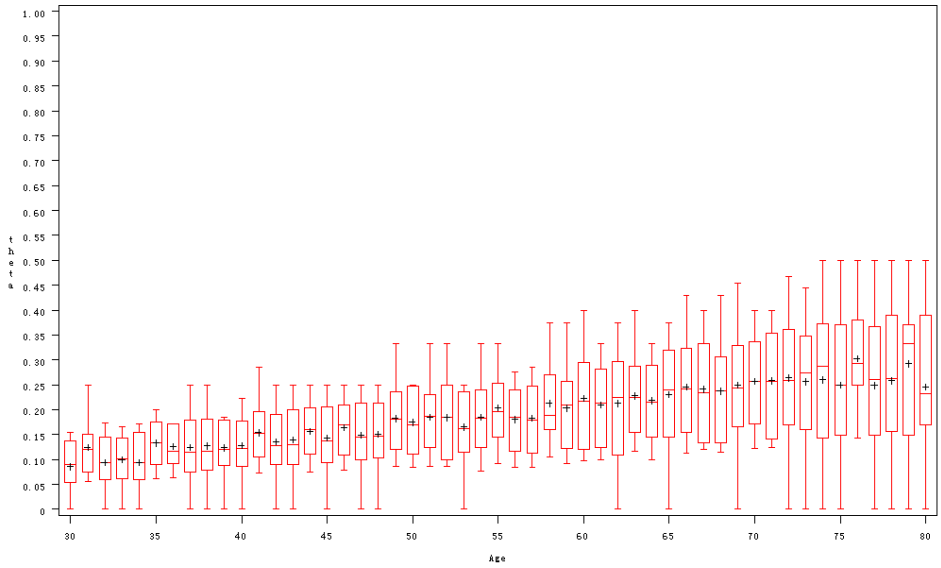

Figure 1.4: Boxplot of the proportions of positive ALS for the 12 domains by age using logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Chapter 2

Bayesian Hierarchical Model for Change point: Individual Domains

Here we build a simple model to find the change point for each of the 12 domains and all the domains combined into one domain (i.e., 13 domains). Let y_j denote the number of adults with positive ALS who are j years old, and let n_j denote the total number of adults that are j years old. Thus,

$$y_j \mid \theta_j \overset{ind}{\sim} \text{Binomial}(n_j, \theta_j), \qquad j = 30, \ldots, 80, \tag{1}$$

where θ_j is the probability of positive ALS. Since we are interested in ALS among adults, we only use the data from people whose ages are between 30 and 80. From Figures 1 & 2, we observe that the θ_j's are similar for the early age groups, and then there is a point of onset, k, where θ_j's tend to get larger. Thus, we can assume that the prior distributions for all θ_j are conjugate *Beta* distributions, but those before the onset and those after the onset have different parameters. Thus,

$$\begin{aligned} \theta_j \mid \mu_1, \tau &\overset{iid}{\sim} \text{Beta}(\mu_1\tau, (1-\mu_1)\tau) \qquad j = 30, \ldots, k \\ \theta_j \mid \mu_2, \tau &\overset{iid}{\sim} \text{Beta}(\mu_2\tau, (1-\mu_2)\tau) \qquad j = k+1, \ldots, 80. \end{aligned} \tag{2}$$

We believe that the change point occurs between the ages 40 and 70, so the range for k is from 40 to 70. A uniform prior distribution is assumed on k, such that

$$P(k = a_r) = w_r, \quad r = 40, \ldots, 70, \quad a_r = r, \quad w_r = \frac{1}{70 - 40 + 1} = \frac{1}{31}. \tag{3}$$

Then non-informative priors are set to the hyperparameters μ and τ,

$$\begin{aligned} &\mu_1, \mu_2 \overset{iid}{\sim} \text{Uniform}(0, 1) \\ &P(\tau) = \tfrac{1}{(1+\tau)^2}, \qquad \tau \geq 0. \end{aligned} \tag{4}$$

All prior distributions must be proper in this analysis.

Then, using Bayes theorem, the joint posterior density of θ, μ, τ, and k is

$$P(\underset{\sim}{\theta}, \underset{\sim}{\mu}, \tau, k|\underset{\sim}{y}) \quad \propto \quad \frac{1}{31}\frac{1}{(1+\tau)^2}\prod_{j=30}^{80}\binom{n_j}{y_j}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \tag{5}$$

$$\prod_{j=30}^{k}\frac{\theta_j^{\mu_1\tau-1}(1-\theta_j)^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)}\prod_{j=k+1}^{80}\frac{\theta_j^{\mu_2\tau-1}(1-\theta_j)^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)}. \tag{6}$$

It turns out to be convenient to collapse over θ, thus we can obtain the joint posterior of $\underset{\sim}{\mu}$ and τ given k as

$$P(\mu_1, \mu_2, \tau \mid \underset{\sim}{y}, k = a_r) \quad \propto \quad \frac{1}{(1+\tau)^2}\prod_{j=30}^{k}\left[\frac{B\left(y_j + \mu_1\tau, (n_j - y_j) + (1-\mu_1)\tau\right)}{B\left(\mu_1\tau, (1-\mu_1)\tau\right)}\right] \tag{7}$$

$$\prod_{j=k+1}^{80}\left[\frac{B\left(y_j + \mu_2\tau, (n_j - y_j) + (1-\mu_2)\tau\right)}{B\left(\mu_2\tau, (1-\mu_2)\tau\right)}\right]. \tag{8}$$

A griddy Gibbs sampler is used to draw samples $(\mu_1^{(h)}, \tau^{(h)})$ and $(\mu_2^{(h)}, \tau^{(h)})$ from this distribution, where $h = 1, \ldots, M, M \approx 1000$. We use a grid of 50 on (0,1) for μ_1 and μ_2, and a grid of 100 on (0,1) for τ. We run 1000 iterations and "burn in" the first 100. Based on these, the conditional density of θ can be obtained.

$$\begin{aligned} &\theta_j \mid \mu_1, \tau, \underset{\sim}{y}, k = a_r \overset{iid}{\sim} \text{Beta}\left(y_j + \mu_1\tau, (n_j - y_j) + (1-\mu_1)\tau\right), \quad j = 30, \ldots, k \\ &\theta_j \mid \mu, \tau, \underset{\sim}{y}, k = a_r \overset{iid}{\sim} \text{Beta}\left(y_j + \mu_2\tau, (n_j - y_j) + (1-\mu_2)\tau\right), \quad j = k+1, \ldots, 80. \end{aligned} \tag{9}$$

Then we can construct $\theta^{(h)}$ and make inference about it corresponds to $(\mu^{(h)}, \tau^{(h)})$ that are drawn from the griddy Gibbs sampler.

The posterior distribution of k is

$$P(k = a_r \mid y) = \frac{P(k = a_r)P(y \mid k = a_r)}{\sum_{s=40}^{70} P(k = a_s)P(y \mid k = a_s)} = \frac{P(y \mid k = a_r)}{\sum_{s=40}^{70} P(y \mid k = a_s)}, \quad r = 40, \ldots, 70, \tag{10}$$

where

$$\begin{aligned} P(\underset{\sim}{y} \mid k = a_r) &= \prod_{j=30}^{80} \binom{n_j}{y_j} \int_{\underset{\sim}{\mu}} \int_{\tau} \frac{1}{(1+\tau)^2} && (11) \\ & \prod_{j=30}^{k} \left[\frac{B\left(y_j + \mu_1\tau, (n_j - y_j) + (1-\mu_1)\tau\right)}{B\left(\mu_1\tau, (1-\mu_1)\tau\right)} \right] && (12) \\ & \prod_{j=k+1}^{80} \left[\frac{B\left(y_j + \mu_2\tau, (n_j - y_j) + (1-\mu_2)\tau\right)}{B\left(\mu_2\tau, (1-\mu_2)\tau\right)} \right] d\underset{\sim}{\mu} d\tau. && (13) \end{aligned}$$

Thus we need to compute $P(\underset{\sim}{y} \mid k = a_r)$ for each a_r. The Monte Carlo integration is applied to realize this computation. To compute the integration of a complicated function, we can multiply it with its density function on both the numerator and the denominator. Then the integration can be estimated based on samples drawn from the distribution, which is also called the importance function.

$$\int g(\mu, \tau) d\mu d\tau = \int \frac{g(\mu, \tau)}{f(\mu, \tau)} f(\mu, \tau) d\mu d\tau \approx \frac{1}{M} \sum_{h=1}^{M} \frac{g(\mu, \tau)}{f(\mu, \tau)}. \tag{14}$$

For our model,

$$\begin{aligned} g(\underset{\sim}{\mu}, \tau) &= \prod_{j=30}^{80} \binom{n_j}{y_j} \frac{1}{(1+\tau)^2} \prod_{j=30}^{k} \left[\frac{B\left(y_j + \mu_1\tau, (n_j - y_j) + (1-\mu_1)\tau\right)}{B\left(\mu_1\tau, (1-\mu_1)\tau\right)} \right] && (15) \\ & \prod_{j=k+1}^{80} \left[\frac{B\left(y_j + \mu_2\tau, (n_j - y_j) + (1-\mu_2)\tau\right)}{B\left(\mu_2\tau, (1-\mu_2)\tau\right)} \right]. && (16) \end{aligned}$$

Beta distributions for $\underset{\sim}{\mu}$, and Gamma distributions for τ are used as importance functions.

$$\mu_i \overset{iid}{\sim} \text{Beta}(\nu_i\phi_i, (1-\nu_i)\phi_i) \quad i = 1, 2 \tag{17}$$

$$\tau \sim \text{Gamma}(\alpha, \beta), \tag{18}$$

and so,

$$f(\underset{\sim}{\mu}, \tau) = \prod_{i=1}^{2} \frac{\mu_i^{\nu_i\phi_i - 1}(1-\mu_i)^{(1-\nu_i)\phi_i - 1}}{B(\nu_i\phi_i, (1-\nu_i)\phi_i)} \frac{\beta^\alpha \tau^{\alpha-1} e^{-\beta\tau}}{\Gamma(\alpha)}. \tag{19}$$

Since

$$E(\mu_i) = \nu_i, \qquad Var(\mu_i) = \frac{\nu_i(1-\nu_i)}{1+\phi_i}, \tag{20}$$

then we can solve for ν_i and ϕ_i,

$$\nu_i = \frac{1}{M}\sum_{h=1}^{M} \mu_i^{(h)}, \qquad \phi_i = \frac{\nu_i(1-\nu_i)}{\frac{1}{M-1}\sum_{h=1}^{M}(\mu_i^{(h)} - \frac{1}{M}\sum_{h=1}^{M}\mu_i^{(h)})^2}. \tag{21}$$

Similarly, for τ, since

$$E(\tau) = \frac{\alpha}{\beta}, \qquad Var(\tau) = \frac{\alpha}{\beta^2}, \tag{22}$$

we can solve for α and β,

$$\alpha = \beta \frac{1}{M}\sum_{h=1}^{M} \tau^{(h)} \tag{23}$$

$$\beta = \frac{\frac{1}{M}\sum_{h=1}^{M}\tau^{(h)}}{\frac{1}{M-1}\sum_{h=1}^{M}(\tau^{(h)} - \frac{1}{M}\sum_{h=1}^{M}\tau^{(h)})^2}. \tag{24}$$

$(\mu_i^{(h)}, \tau^{(h)}), h = 1, \ldots, M, M \approx 1000$ are obtained from the Gibbs sampler.

Now we can get inferences about the change point k. Results for each domain are presented in Tables 2.1 to 2.4. Domain 13 is using the overall data. From the table we can see that, for the unweighted data, the onset for the overall population occurs around 47 to 50, most likely to be 48 and 49 for case 1 and case 2 respectively, where in case 1 we only inculde "Limited in major activity" as positive ALS and in case 2, both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS. For the weighted data, the onset for the overall population occurs around 56 to 59, and age 59 is most likely to be the onset

for both case. For each individual domains, including the condition "limited in kind/amount of major activity" does not necessarily produce a later onset. For most domains, especially when the weighted data are used, the distributions for k spread out the whole range from 40 to 70. For Table 2.1, the onset probably occurs between 40 to 58, though domains 1, 2, and 4 have a later onset around age 69. For Table 2.2, the onset probably occurs between age 40 to 56, though domains 2 and 12 have later onsets around 69 and 62 respectively. For the weighted data as presented in Tables 2.2 and 2.3, the onset probably occurs between age 45 to 60, though domains 10 and 12 have a later onset around 70. After a reexamination of the data, we find that this may be due to the lack of data for domains associated with race for low and high education.

Based on the distribution of k, we can make inferences about the parameter $\underset{\sim}{\theta}$. Since

$$P(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau, k|\underset{\sim}{y}) = P(\underset{\sim}{\theta}|\mu_1, \mu_2, \tau, k, \underset{\sim}{y})P(\mu_1, \mu_2, \tau|k, \underset{\sim}{y})P(k|\underset{\sim}{y}), \tag{25}$$

we can first draw k from $P(k|\underset{\sim}{y})$, and for each given k, draw μ_1, μ_2, τ from $P(\mu_1, \mu_2, \tau|k, \underset{\sim}{y})$. Repeat this process for 1000 times, then we can draw $\underset{\sim}{\theta}$ based on these.

$$\begin{aligned} \theta_j \mid \mu_1, \tau, k, \underset{\sim}{y} &\overset{iid}{\sim} \text{Beta}\left(y_j + \mu_1\tau, n_j - y_j + (1-\mu_1)\tau\right) \qquad j = 30, \ldots, k \\ \theta_j \mid \mu_2, \tau, k, \underset{\sim}{y} &\overset{iid}{\sim} \text{Beta}\left(y_j + \mu_2\tau, n_j - y_j + (1-\mu_2)\tau\right) \qquad j = k+1, \ldots, 80. \end{aligned} \tag{26}$$

Figures 2.1 to 2.4 are plots for the posterior mean and 95% pointwise credible bands for $\underset{\sim}{\theta}$ with observed data for each domain. Generally speaking, the weighted data are much smoother than the unweighted data. Compared with the case where only "Limited in major activity" is considered as positive ALS, $\underset{\sim}{\theta}$s are higher when both "Limited in major activity" and "Limited in king/amound of major activity" are considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.038	.003	.102	.028	.058	.074	.037	.001	.001	.043	.013	.007	.009
41	.034	.003	.317	.027	.038	.021	.023	.003	.002	.027	.031	.004	.011
42	.039	.005	.022	.028	**.140**	.073	.035	.005	.003	.022	**.526**	.003	.030
43	.060	.004	.080	.031	.099	**.115**	.096	.002	.006	.020	.142	.003	.050
44	.070	.009	**.377**	.040	.116	.108	.061	.001	.012	.015	.015	.004	.042
45	.092	.006	.052	.028	.136	.066	.111	.002	.069	.034	.117	.008	.104
46	.131	.004	.023	.034	.066	.069	.057	.005	.019	.028	.007	.024	.067
47	.112	.003	.014	.039	.060	.064	.058	.011	.122	.018	.015	.013	.103
48	.050	.003	.005	.043	.082	.110	.073	.096	**.273**	.014	.010	.014	**.170**
49	.019	.003	.003	.029	.055	.019	**.178**	**.491**	.148	.016	.008	.044	.152
50	.010	.002	.002	.035	.028	.023	.142	.336	.044	.017	.028	**.135**	.118
51	.012	.003	.001	.027	.017	.047	.083	.029	.119	.028	.018	.118	.075
52	.012	.003	.000	.023	.030	.037	.014	.010	.043	.059	.015	.038	.032
53	.006	.003	.001	.022	.010	.022	.008	.003	.021	.069	.008	.081	.014
54	.006	.004	.000	.022	.010	.038	.010	.001	.024	.045	.023	.034	.012
55	.004	.003	.000	.021	.007	.051	.003	.001	.030	.029	.013	.042	.005
56	.004	.004	.000	.021	.009	.037	.005	.000	.063	.051	.008	.021	.004
57	.002	.006	.000	.022	.002	.011	.002	.000	.001	**.098**	.001	.020	.001
58	.002	.004	.000	.021	.001	.003	.001	.000	.000	.044	.002	.042	.000
59	.002	.004	.000	.021	.001	.001	.000	.000	.000	.038	.000	.113	.000
60	.001	.004	.000	.020	.001	.001	.000	.000	.000	.018	.001	.072	.000
61	.001	.004	.000	.021	.000	.000	.000	.000	.000	.029	.000	.035	.000
62	.001	.004	.000	.021	.000	.000	.000	.000	.000	.052	.000	.052	.000
63	.002	.005	.000	.023	.000	.000	.000	.000	.000	.050	.000	.010	.000
64	.002	.015	.000	.021	.000	.000	.000	.000	.000	.033	.000	.019	.000
65	.004	.015	.000	.028	.000	.000	.000	.000	.000	.023	.000	.011	.000
66	.013	.033	.000	.026	.001	.001	.000	.000	.000	.011	.000	.008	.000
67	.013	.047	.000	.033	.002	.001	.000	.000	.000	.012	.000	.006	.000
68	.023	.106	.000	.074	.007	.001	.000	.000	.000	.015	.000	.009	.000
69	**.178**	**.533**	.000	**.115**	.016	.004	.000	.000	.000	.017	.000	.004	.000
70	.057	.157	.000	.056	.005	.002	.000	.001	.000	.026	.000	.005	.000

Table 2.1: Distributions for the change k for the 13 domains using the unweighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.073	.032	.040	.037	.053	.030	**.267**	.002	.000	.003	.040	.063	.002
41	.065	.035	**.586**	.040	.037	.013	.056	.006	.001	.004	.020	.015	.003
42	.089	.027	.005	.045	.126	.043	.041	.010	.004	.004	.080	.008	.007
43	.095	.026	.127	.031	.209	.071	.078	.017	.003	.007	.025	.005	.015
44	.155	.030	.124	.075	**.149**	.090	.031	.004	.053	.008	.016	.002	.015
45	**.220**	.026	.073	.036	.104	.084	.050	.006	**.534**	.034	**.259**	.013	.085
46	.135	.024	.023	.074	.066	.090	.028	.038	.045	.033	.064	.022	.067
47	.092	.023	.009	.060	.077	.072	.045	.053	.093	.048	.074	.016	.135
48	.041	.023	.002	**.152**	.046	.076	.043	**.685**	.024	.037	.064	.016	.174
49	.009	.023	.005	.081	.026	.026	.203	.087	.034	.009	.090	.022	**.221**
50	.003	.029	.004	.056	.021	.020	.083	.070	.037	.010	.102	.034	.147
51	.004	.025	.001	.027	.013	.064	.038	.017	.124	.018	.041	.030	.076
52	.004	.023	.000	.027	.038	.042	.007	.002	.030	.120	.033	.006	.027
53	.002	.024	.000	.024	.014	.034	.004	.001	.012	**.427**	.010	.019	.012
54	.002	.022	.000	.032	.007	.058	.006	.001	.003	.061	.033	.004	.008
55	.001	.024	.000	.016	.006	**.091**	.002	.001	.002	.007	.029	.009	.004
56	.001	.021	.000	.011	.003	.067	.009	.000	.002	.010	.007	.008	.002
57	.001	.020	.000	.013	.001	.013	.003	.000	.000	.035	.002	.005	.000
58	.001	.020	.000	.016	.001	.005	.004	.000	.000	.025	.003	.017	.000
59	.000	.020	.000	.018	.000	.002	.001	.000	.000	.039	.001	.087	.000
60	.000	.020	.000	.018	.000	.001	.001	.000	.000	.011	.005	.095	.000
61	.000	.020	.000	.016	.000	.001	.000	.000	.000	.007	.002	.105	.000
62	.000	.020	.000	.017	.000	.001	.000	.000	.000	.009	.000	**.248**	.000
63	.000	.023	.000	.021	.000	.001	.000	.000	.000	.019	.000	.045	.000
64	.000	.035	.000	.011	.000	.000	.000	.000	.000	.007	.000	.071	.000
65	.000	.037	.000	.008	.000	.000	.000	.000	.000	.003	.000	.011	.000
66	.000	.049	.000	.007	.000	.001	.000	.000	.000	.001	.000	.006	.000
67	.000	.049	.000	.007	.000	.001	.000	.000	.000	.002	.000	.003	.000
68	.000	.063	.000	.007	.000	.001	.000	.000	.000	.002	.000	.008	.000
69	.002	**.130**	.000	.008	.000	.003	.000	.000	.000	.001	.000	.003	.000
70	.002	.059	.000	.008	.000	.002	.000	.000	.000	.004	.000	.003	.000

Table 2.2: Distributions for the change point k for the 13 domains using the unweighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.008	.027	.001	.002	.001	.032	.000	.011	.015	.029	.014	.023	.000
41	.010	.021	.001	.002	.002	.030	.001	.013	.018	.027	.017	.023	.000
42	.011	.040	.002	.005	.003	.028	.001	.018	.021	.028	.019	.021	.000
43	.013	**.082**	.003	.011	.005	.032	.002	.022	.026	.030	.020	.020	.000
44	.018	.038	.004	.008	.008	.034	.005	.025	.031	.028	.023	.021	.000
45	.020	.060	.006	.011	.013	.029	.008	.026	.040	.025	.028	.021	.000
46	.025	.043	.008	.025	.018	.033	.012	.033	.048	.026	.029	.020	.000
47	.029	.027	.013	.038	.021	.034	.018	.037	.055	.024	.034	.020	.000
48	.029	.042	.019	.055	.031	.034	.023	.036	**.061**	.024	.038	.021	.000
49	.036	.034	.028	.110	.035	.036	.028	.036	.056	.023	.035	.021	.002
50	.051	.029	.027	.069	.051	.042	.033	.040	.052	.025	.033	.022	.005
51	**.073**	.036	.023	.046	.071	**.049**	.036	.042	.053	.027	.036	.022	.012
52	.065	.031	.032	.049	.076	.043	.045	.040	.049	.030	.035	.023	.024
53	.056	.026	.046	.077	.091	.043	.065	.043	.051	.028	.032	.024	.053
54	.053	.034	.058	.068	.081	.044	.079	.043	.043	.026	.036	.025	.080
55	.050	.047	.057	**.135**	**.094**	.038	.077	.043	.036	.026	.033	.026	.079
56	.064	.044	.062	.122	.076	.045	.084	.036	.038	.026	.040	.028	.125
57	.061	.051	.059	.057	.063	.041	.093	.036	.040	.027	.043	.031	.152
58	.061	.039	.066	.036	.050	.034	**.106**	.038	.041	.028	.050	.035	.135
59	.057	.037	.090	.029	.041	.037	.098	.047	.040	.028	**.062**	.042	**.234**
60	.054	.045	**.091**	.015	.047	.026	.079	.046	.028	.030	.051	.038	.067
61	.038	.031	.078	.009	.034	.028	.053	**.055**	.029	.032	.044	.047	.026
62	.028	.032	.054	.005	.027	.029	.026	.051	.028	.036	.032	.043	.005
63	.026	.033	.041	.003	.024	.022	.013	.041	.021	.041	.023	.040	.001
64	.025	.022	.036	.003	.015	.024	.007	.042	.018	.039	.027	.054	.000
65	.015	.012	.024	.002	.010	.024	.004	.031	.016	.039	.029	.050	.000
66	.008	.010	.023	.002	.006	.022	.003	.027	.012	.041	.028	.046	.000
67	.006	.008	.019	.002	.003	.023	.001	.017	.009	.048	.029	.043	.000
68	.005	.009	.014	.001	.001	.023	.001	.011	.009	.049	.026	.043	.000
69	.003	.006	.011	.001	.001	.021	.000	.007	.008	.053	.026	**.055**	.000
70	.002	.005	.005	.001	.000	.019	.000	.007	.006	**.058**	.028	.053	.000

Table 2.3: Distributions for the change point k for the 13 domains using the logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40	.001	.007	.001	.009	.000	.018	.000	.015	.004	.037	.004	.034	.000
41	.002	.007	.001	.007	.000	.022	.000	.021	.006	.040	.006	.031	.000
42	.003	.015	.002	.007	.001	.028	.000	.024	.009	.036	.009	.032	.000
43	.005	.034	.003	.007	.002	.027	.000	.028	.013	.033	.011	.032	.000
44	.007	.027	.006	.010	.005	.025	.001	.037	.019	.030	.017	.034	.000
45	.010	.046	.009	.011	.010	.027	.002	.043	.024	.029	.021	.035	.000
46	.015	.043	.013	.012	.016	.031	.004	.052	.030	.031	.028	.033	.000
47	.016	.040	.020	.020	.023	.040	.008	.050	.040	.030	.036	.031	.000
48	.023	.068	.025	.025	.033	.044	.016	.049	.042	.029	.045	.027	.000
49	.025	.068	.032	.022	.043	.045	.029	.051	.041	.028	.047	.027	.000
50	.025	**.068**	.035	.025	.065	.054	.049	.048	.050	.030	.060	.027	.001
51	.034	.045	.056	.028	.096	.049	.066	.048	.062	.029	.065	.026	.003
52	.037	.047	.059	.023	**.098**	.049	.079	.055	.062	.032	**.073**	.028	.008
53	.053	.055	.065	.039	.096	.040	.104	**.057**	.062	.028	.056	.027	.023
54	.069	.041	**.086**	.058	.093	.044	.111	.045	.060	.025	.051	.029	.057
55	.091	.032	.076	.053	.081	.042	.102	.047	.057	.023	.055	.028	.060
56	.075	.036	.081	.083	.081	.051	.131	.044	.059	.025	.051	.028	.142
57	.087	.052	.066	**.093**	.071	**.065**	.096	.040	.067	.028	.054	.028	.304
58	.087	.055	.061	.061	.055	.050	.071	.051	**.069**	.026	.044	.027	.168
59	**.113**	.034	.060	.075	.044	.038	.052	.050	.059	.028	.043	.028	**.187**
60	.073	.024	.045	.052	.036	.032	.032	.036	.042	.030	.046	.029	.031
61	.055	.022	.045	.065	.020	.033	.021	.032	.036	.033	.035	.030	.013
62	.032	.015	.043	.050	.012	.031	.013	.021	.029	.036	.037	.031	.002
63	.021	.019	.035	.032	.009	.023	.006	.015	.020	.034	.029	.032	.000
64	.019	.018	.034	.035	.005	.024	.004	.013	.012	.034	.025	.034	.000
65	.011	.014	.018	.031	.002	.016	.002	.010	.008	.034	.016	.036	.000
66	.006	.014	.011	.022	.001	.012	.001	.008	.006	.034	.013	.037	.000
67	.003	.015	.006	.019	.000	.012	.000	.004	.005	.036	.008	.039	.000
68	.002	.020	.003	.014	.000	.011	.000	.003	.004	.040	.007	.040	.000
69	.001	.011	.002	.007	.000	.010	.000	.001	.002	.043	.005	.045	.000
70	.001	.010	.001	.007	.000	.008	.000	.001	.001	**.048**	.004	**.054**	.000

Table 2.4: Distributions for the change point k for the 13 domains using the logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Figure 2.1: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where only "Limited in major activity " is considered as positive ALS.

Figure 2.2: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where both “Limited in major activity” and “Limited in kind/amount of major activity ” are considered as positive ALS.

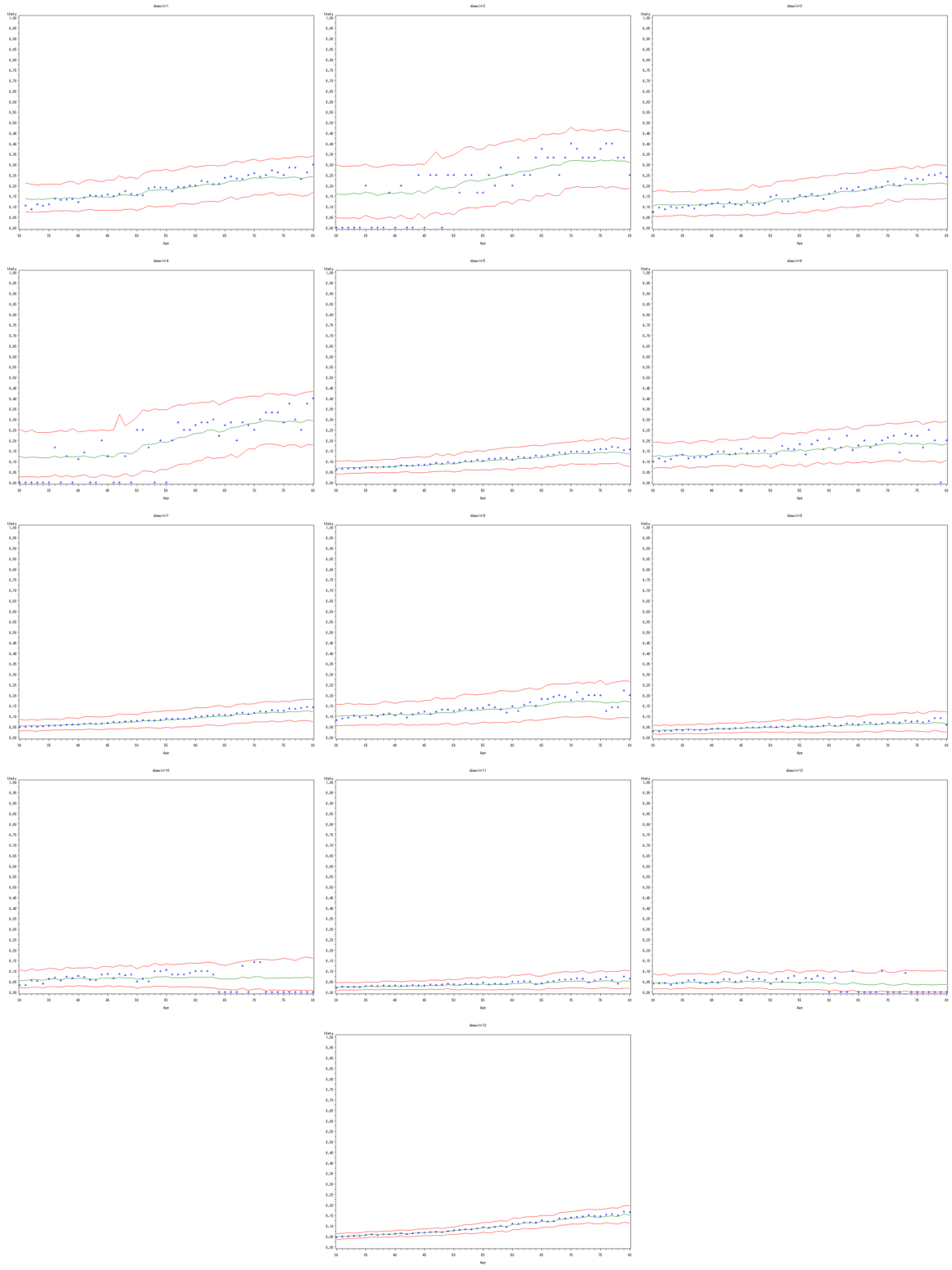

Figure 2.3: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where only "Limited in major activity " is considered as positive ALS.

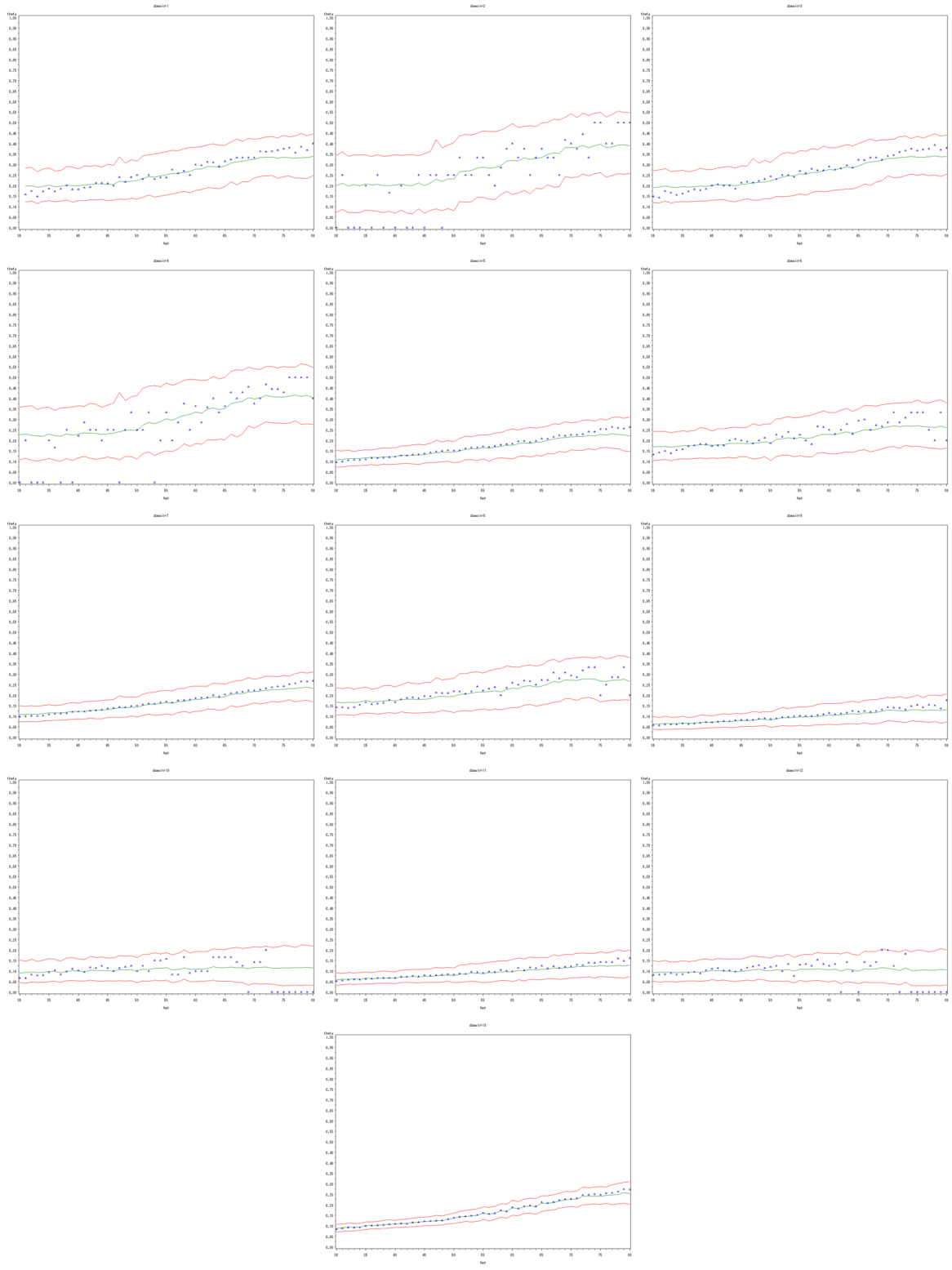

Figure 2.4: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where both “Limited in major activity” and “Limited in kind/amount of major activity ” are considered as positive ALS.

Chapter 3

Reversible Jump Model

The procedure described in Chapter 2 is computational intensive, so we wonder how to reduce the computation effort.Also, when there are many competing models with different parameter spaces (e.g., dimensions), there is uncertainty about the model itself creating a parameter which indexes the model. To solve this problem, Green [7] proposed a reversible Markov chain samplers that jump between parameter subspaces of differing dimensionality.

There are four components for this process. In our problem, first, there should be individual models of y_j given parameters θ_k and priors $p(k = a_r)$, where $a_r = 40, \ldots, 70$ for the 31 distinct models. The individual model is

$$\begin{aligned} y_j \mid \theta_j, k &\overset{ind}{\sim} \text{Binomial}(n_j, \theta_j), \qquad j = 30, \ldots, 80 \\ \theta_j \mid \mu_1, \tau &\overset{iid}{\sim} \text{Beta}(\mu_1\tau, (1-\mu_1)\tau) \qquad j = 30, \ldots, k \\ \theta_j \mid \mu_2, \tau &\overset{iid}{\sim} \text{Beta}(\mu_2\tau, (1-\mu_2)\tau) \qquad j = k+1, \ldots, 80. \end{aligned} \tag{1}$$

Second, we should have prior probabilities for k, a discrete uniform distribution on [40,70].

$$p(k = a_r) = w_{a_r} = \frac{1}{31}, \qquad a_r = 40, \ldots, 70. \tag{2}$$

Then the posterior density for k can be written as

$$p(k \mid \underset{\sim}{y}) = \frac{w_{k=a_r}\pi(\underset{\sim}{\theta}_{k=a_r} \mid k = a_r) f(\underset{\sim}{y} \mid \underset{\sim}{\theta}_{k=a_r}, k = a_r)}{\sum_{k=40}^{70} w_k \int \pi(\underset{\sim}{\theta}_k \mid k) f(\underset{\sim}{y} \mid \underset{\sim}{\theta}_k, k) d\underset{\sim}{\theta}_k}, \quad a_r = 40, \ldots, 70, \tag{3}$$

where

$$\pi(\underset{\sim}{\theta}_k \mid k) f(\underset{\sim}{y} \mid \underset{\sim}{\theta}_k, k) = \frac{1}{31}\frac{1}{(1+\tau)^2}\prod_{j=30}^{80}\binom{n_j}{y_j}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \tag{4}$$

$$\prod_{j=30}^{k}\frac{\theta_j^{\mu_1\tau-1}(1-\theta_j)^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)}\prod_{j=k+1}^{80}\frac{\theta_j^{\mu_2\tau-1}(1-\theta_j)^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)}, \tag{5}$$

and the selected model is,

$$p(k=a_r^* \mid \underset{\sim}{y}) = \max_{a_r=40,\ldots,70} p(k=a_r \mid \underset{\sim}{y}). \tag{6}$$

To predict $\underset{\sim}{y}^{(p)}$, we can use

$$\pi(\underset{\sim}{y}^{(p)} \mid \underset{\sim}{y}) = \sum_{k=40}^{70}\pi(y^{(p)} \mid y,k)p(k \mid \underset{\sim}{y}). \tag{7}$$

Third, there should be transition probabilities between consecutive models. These are not a part of the model specifications, but are chosen to provide good moves. The range for the change point is between 40 and 70. We define a birth as the change point increases by 1, and a death as the change point decreases by 1. At age 40, only a birth is allowed, and at age 70, only death is allowed. At all other points, it can either have a birth or a death, or stay where it is. We assign the transition probabilities as follows,

$$\pi_{40,41} = 0.99 \qquad \pi_{40,40} = 0.01 \tag{8}$$

$$\pi_{70,69} = 0.99 \qquad \pi_{70,70} = 0.01 \tag{9}$$

$$\pi_{a_r,a_r+1} = 0.45 \qquad \pi_{a_r,a_r} = 0.10 \qquad \pi_{40,41} = 0.99\pi_{a_r,a_r-1} = 0.45. \tag{10}$$

Finally, we need some dimension-matching condition. Since $\dim(\underset{\sim}{\theta}_{a_r+1}) > \dim(\underset{\sim}{\theta}_{a_r})$, we introduce one latent variables into model $k = a_r$ to match the dimension of the model $k = a_r + 1$, say u. Then we need a bijection:

$$\underset{\sim}{\theta}_{a_r+1} = A\begin{pmatrix}\underset{\sim}{\theta}_{a_r}\\ u\end{pmatrix}; \qquad \begin{pmatrix}\underset{\sim}{\theta}_{a_r}\\ u\end{pmatrix} = A^{-1}\underset{\sim}{\theta}_{a_r+1}. \tag{11}$$

In our problem, the Jacobian of the bijection is 1.

Now we can use the griddy sampler as that in Chapter 2, and draw samples for μ_1, μ_2, and τ. Using these, we can obtain the joint posterior distribution for $\underset{\sim}{\theta}$ and k

$$p(\underset{\sim}{\theta}_{a_r}, \underset{\sim}{\mu}, \tau, k = a_r \mid \underset{\sim}{y}) = \frac{1}{31}\frac{1}{(1+\tau)^2}\prod_{j=30}^{80}\binom{n_j}{y_j}\theta_j^{y_j}(1-\theta_j)^{n_j-y_j} \tag{12}$$

$$\prod_{j=30}^{k}\frac{\theta_j^{\mu_1\tau-1}(1-\theta_j)^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)}\prod_{j=k+1}^{80}\frac{\theta_j^{\mu_2\tau-1}(1-\theta_j)^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)}. \tag{13}$$

Based on these posterior densities when $k = a_r$ and $k = a_r + 1$, we can compute the acceptance probability of moving from model $k = a_r$ to the model $k = a_r + 1$.

$$R = \frac{p(\underset{\sim}{\theta}_{a_r+1}, \underset{\sim}{\mu}, \tau, k = a_r + 1 \mid \underset{\sim}{y})\pi_{a_r+1,a_r}q(u)}{p(\underset{\sim}{\theta}_{a_r}, \underset{\sim}{\mu}, \tau, k = a_r \mid \underset{\sim}{y})\pi_{a_r,a_r+1}}\left|\frac{d(\underset{\sim}{\theta}_{a_r+1}, u)}{d\underset{\sim}{\theta}_{a_r}}\right|. \tag{14}$$

The acceptance probability for the move of the change point k from a_r to a_r+1 is $\min(1, R)$. We draw a random number from 0 and 1. If this number is smaller than the acceptance probability, then change point increases by 1 and k moves from a_r to a_r+1. So we have a birth. Similarly, we also compute the posterior density when the change point occurs at $a_r - 1$, and

$$R^{-1} = \frac{p(\underset{\sim}{\theta}_{a_r-1}, \underset{\sim}{\mu}, \tau, k = a_r - 1 \mid \underset{\sim}{y})\pi_{a_r-1,a_r}q(u)}{p(\underset{\sim}{\theta}_{a_r}, \underset{\sim}{\mu}, \tau, k = a_r \mid \underset{\sim}{y})\pi_{a_r,a_r-1}}\left|\frac{d(\underset{\sim}{\theta}_{a_r-1}, u)}{d\underset{\sim}{\theta}_{a_r}}\right|. \tag{15}$$

The acceptance probability for the move from a_r to $a_r - 1$ is $\min(1, R^{-1})$. Draw a random number from 0 and 1, if it is smaller than the acceptance probability, we have a death and k moves from a_r to $a_r - 1$. The change point decreases by 1.

Thus the change point k jumps between 40 and 70. The distributions for k are presented in Tables 3.1 to 3.4. Inferences about $\underset{\sim}{\theta}$ can also be obtained and the posterior mean and 95% credible bands are ploted with observed data in Figures 3.1 to 3.4.

Compared with Tables 2.1 to 2.4, the distributions for k using the reversible jump model. For the onset of the overall population, the unweighted data produce different result for case 1 and case 2. The onset when only "Limited in major activity" is considered as positive ALS occurs around 69, while the onset for case 2 where "Limited in kind/amount of activity limitation" is also considered as ALS, the onset occurs around 49, which is similar to the result in Chapter

2. For the weighted data, case 1 and case 2 have similar results. The onset occurs between 54 to 59, and 56 is most likely to be the onset. This also corresponds to the results in Chapter 2. Comparing Table 3.1 to Table 2.1, the distributions of k using the reversible jump model center around age 68 to 70. Domains 2, 4, 6, 10, and 12 have probability 1 at age 70. Domains 9 and 11 are more spread out. For Table 3.2, onsets are also around 68 to 70. Domains 2, 10, and 12 have probability 1 at age 70. Tables 3.3 and 3.4 using the weighted data is more spread out than the previous 2 tables using the unweighted data. However, the onsets are still later than those in Chapter 2, and domains 10 and 12 have probability 1 at age 70. This means that the reversible jump model gets stuck at the border line, which will interfere with inference.

Thus, in future, we would not use the reversible jump sampler for the change point problem. Simply, it gets stuck at the boundaries of the parameter space, making inference difficult or impossible.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40									.013				
41									.017				
42									.015				
43									.011				
44									.024				
45									.021				
46									.016				
47									.048				
48									.047		.004		
49									.035		.001		
50									.046		.014		
51									.075		.007		
52									.055		.007		
53									.045		.005		
54									.074		.006		
55									**.143**		.008		
56									.141		.004		
57									.036		.002		
58									.024		.003		
59									.028		.007		
60									.009		.007		
61									.012		.003		
62									.007		.009		
63									.007		.005		
64									.003		.051		
65									.009		.077		.001
66			.010				.002		.007		.078		.003
67	.001		.022		.001		.006	.001	.011		.170		.020
68	.107		.280		**.225**		.187	.112	.010		**.221**		.276
69	**.639**		.350		.570		**.415**	.226	.008		.174		**.402**
70	.253	**1.00**	**.338**	**1.00**	.204	**1.00**	.390	**.661**	.003	**1.00**	.137	**1.00**	.298

Table 3.1: Distributions for the change point k for the 13 domains using the reversible jump model and unweighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40			.001										
41			.011										
42			.001						.002				
43			.066						.003				
44			.059						.055				.001
45			.035						.086				.005
46			.010						.041				.016
47			.009						.056				.057
48			.006						.037				.140
49			.006						.047				**.231**
50			.005						.091				.178
51			.002						**.113**				.121
52			.001						.061				.065
53			.013						.030				.046
54			.007						.026		.002		.055
55			.004				.001		.024		.003		.040
56			.000				.006		.017		.001		.029
57			.004				.005		.005		.002		.011
58			.002				.008		.000		.002		.005
59			.002				.004		.003		.016		
60			.000				.019		.000		.065		
61			.000				.011		.001		.057		
62			.000	.001			.020		.001		.016		
63			.002	.001			.011		.002		.030		
64			.001	.002			.019	.002	.002		.041		
65			.001	.002			.011	.004	.004		.071		
66			.030	.013			.011	.011	.015		.045		
67	.001		.097	.048	.003	.003	.048	.020	.052		.128		
68	.136		**.257**	.317	.157	.166	**.298**	.302	.091		**.245**		
69	**.486**		.211	**.327**	**.439**	**.534**	.288	**.349**	.079		.154		
70	.377	**1.00**	.157	.289	.401	.297	.240	.312	.056	**1.00**	.122	**1.00**	

Table 3.2: Distributions for the change point k for the 13 domains using the reversible jump model and unweighted data, where both "Limited in major activity" and "Limited in kind/amout of major activity" are considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40													
41													
42													
43													
44													
45													
46		.001		.001	.001								
47	.001	.000		.002	.000								
48	.000	.006	.002	.006	.000								
49	.001	.003	.002	.022	.001								
50	.000	.001	.001	.009	.001								.007
51	.001	.004	.000	.009	.004		.002						.013
52	.003	.008	.001	.008	.004		.005						.030
53	.000	.006	.005	.027	.007		.001						.074
54	.010	.017	.005	.033	.008		.012						.131
55	.003	.023	.010	**.112**	.008		.013						.134
56	.009	.021	.017	.097	.005		.032						**.183**
57	.025	.037	.025	.052	.015		.063				.001		.161
58	.024	.018	.030	.044	.014		.086				.000		.096
59	.037	.036	.060	.036	.029		.124	.002			.002		.111
60	.059	.035	.073	.043	.033		**.129**	.003			.007		.039
61	.051	.032	.078	.031	.057		.115	.005	.002		.005		.014
62	.072	.049	.055	.028	.055	.001	.112	.008	.003		.008		.005
63	.075	.077	.064	.027	.100	.001	.088	.014	.001		.027		.000
64	.112	.066	.073	.041	.107	.006	.047	.039	.008		.054		.002
65	.072	.058	.076	.030	.129	.005	.051	.039	.026		.077		
66	.079	.079	.101	.047	**.135**	.026	.039	.088	.051		.128		
67	.102	.086	**.118**	.080	.106	.087	.038	.166	.101		.168		
68	**.125**	**.170**	.110	.108	.111	**.325**	.017	**.261**	**.336**		**.259**		
69	.091	.104	.073	.067	.048	.303	.018	.206	.283		.154		
70	.048	.063	.021	.040	.022	.246	.008	.169	.189	**1.00**	.110	**1.00**	

Table 3.3: Distributions for the change point k for the 13 domains using the reversible jump model and logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13
40													
41													
42													
43		.001	.002										
44		.000	.002				.001						
45		.001	.001				.001						
46	.001	.001	.002				.000						
47	.000	.001	.000		.001		.000						
48	.001	.002	.007		.002		.004						.001
49	.000	.001	.006		.004		.009						.006
50	.000	.002	.003	.001	.003		.016						.008
51	.003	.000	.005	.000	.009		.022				.001		.017
52	.003	.003	.010	.002	.010		.031	.001	.002		.002		.048
53	.007	.001	.014	.003	.013		.046	.000	.000		.002		.074
54	.017	.005	.021	.005	.014	.001	.042	.000	.001		.001		.099
55	.026	.002	.028	.016	.029	.000	.045	.001	.002		.002		.106
56	.031	.000	.035	.032	.050	.001	.088	.003	.003		.002		**.162**
57	.067	.012	.042	.042	.069	.000	.075	.002	.008		.005		.155
58	.066	.005	.041	.033	.058	.001	.085	.002	.004		.004		.117
59	**.111**	.008	.064	.048	.079	.003	.090	.004	.009		.014		.118
60	.093	.010	.063	.053	.064	.001	.068	.006	.013		.021		.054
61	.094	.011	.089	.068	.068	.002	**.084**	.020	.027		.025		.024
62	.089	.015	.089	.054	.085	.009	.066	.025	.038		.035		.004
63	.054	.042	.115	.051	**.101**	.012	.049	.035	.038		.028		.005
64	.073	.038	**.118**	.077	.072	.025	.062	.063	.050		.055	.001	.001
65	.073	.039	.071	.083	.084	.031	.043	.086	.094		.084	.002	.000
66	.062	.090	.061	.084	.086	.040	.026	.137	.097		.105	.002	.001
67	.045	.172	.047	**.128**	.057	.128	.017	.158	.178		.137	.010	
68	.036	**.301**	.034	.115	.025	**.289**	.021	**.225**	**.234**		**.221**	.105	
69	.034	.157	.021	.055	.014	.263	.006	.148	.130		.164	.185	
70	.014	.080	.009	.050	.003	.194	.003	.084	.072	**1.00**	.092	**.695**	

Table 3.4: Distributions for the change point k for the 13 domains using the reversible jump model and logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Figure 3.1: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where only "Limited in major activity " is considered as positive ALS.

Figure 3.2: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where both “Limited in major activity” and “Limited in kind/amount of major activity ” are considered as positive ALS.

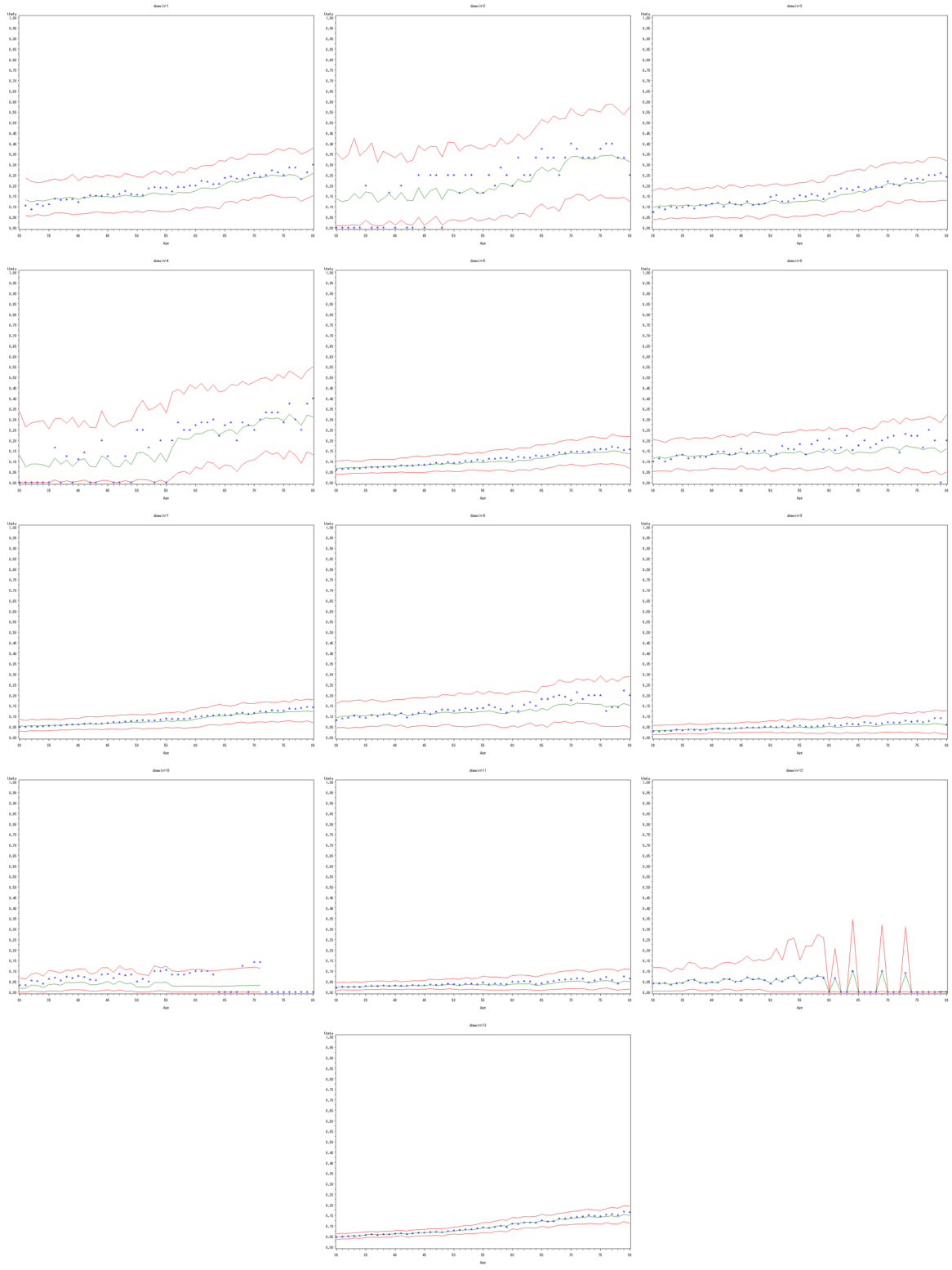

Figure 3.3: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

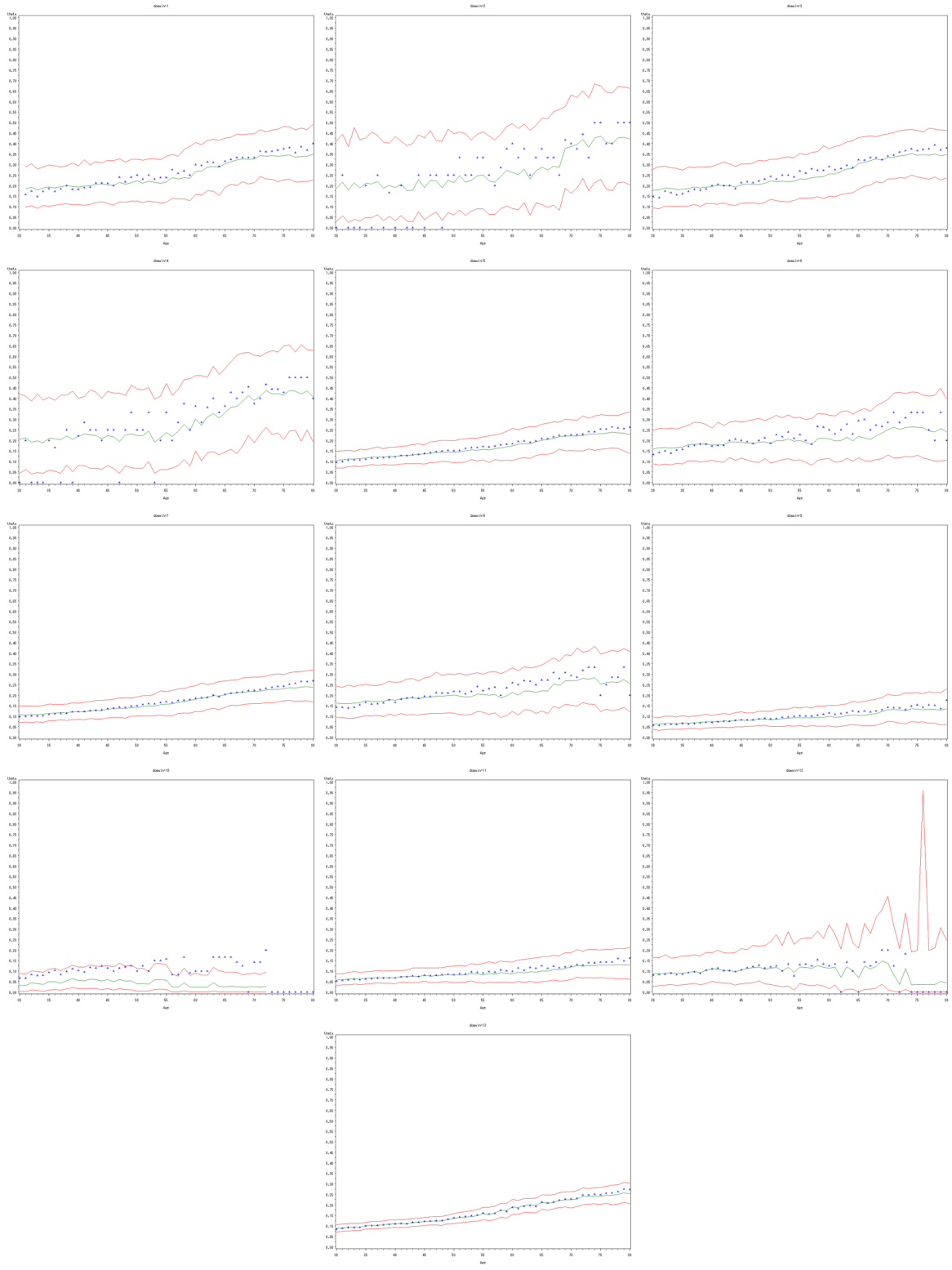

Figure 3.4: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Chapter 4

Bayesian Hierarchical Model: Pooling the Domains

In this chapter we model heterogeneity among the 12 domains. We assume that the parameters of the Binomial distribution follow a common stochastic process.This allows us to pool the domains adaptively (i.e., according to the sample size). This comes naturally under small arear estimation. This chapter has two parts. In the first part, we assume that all domains have the same change point (an unrealistic situation), and in the second part we eliminate this assumption to have different change points for the domains, a more realistic approach. The procedure used in the first part is also used in the second part.

4.1 A single change point for all domains

In Chapter 2, we only computed the distributions of the change point k for each domain separately. The results indicate that the distributions are quite spread out and may not be very accurate. In this chapter, we build another Bayesian hierachical model based on the model in Chapter 2, using not only the data from one domain, but from all domains and still maintaining the domains' identity. So we will have a single model containing all the domains. This model involves two steps. In the first step, we build a single model containing all the domains and same as the model in Chapter 2, we assume that the change point k is the same for all domains. In this step, we will obtain 1) the posterior distributions for k using all the data, 2) samples of

μ_1, μ_2 and τ from the posterior distribution for each k, and 3) posterior distribution of $\underset{\sim}{\theta}$.

Like the previous models, y_{ij}'s, the number of adults with positive ALS who are j years old and in the i^{th} domain, follow a Binomial distribution with parameters n_{ij} and θ_{ij}. θ_{ij}'s have Beta distributions, but the parameters are different for those before the change point and those after the change point.

$$y_{ij}|\theta_{ij} \overset{ind}{\sim} \text{Binomial}(n_{ij}, \theta_{ij}) \quad i = 1, \ldots, 12, \quad j = 30, \ldots, 80 \tag{1}$$

$$\theta_{ij} \mid \mu_1, \tau \overset{iid}{\sim} \text{Beta}(\mu_1\tau, (1-\mu_1)\tau) \qquad j = 30, \ldots, k \tag{2}$$

$$\theta_{ij} \mid \mu_2, \tau \overset{iid}{\sim} \text{Beta}(\mu_2\tau, (1-\mu_2)\tau) \qquad j = k+1, \ldots, 80, i = 1, \ldots, 12. \tag{3}$$

Set Uniform prior distributions on μ_1 and μ_2, and noninformative prior distribution on τ, we have

$$\mu_1, \mu_2 \overset{iid}{\sim} \text{Uniform}(0, 1) \tag{4}$$

$$p(\tau) = \frac{1}{(1+\tau)^2}. \tag{5}$$

Also, the prior distribution for k is also uniform on $(40, 70)$,

$$p(k) = \frac{1}{31}, \qquad k = 40, \ldots, 70. \tag{6}$$

Then the joint posterior distribution for all the parameters will be

$$\begin{aligned} p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau, k|\underset{\sim}{y}) \quad = \quad & \frac{1}{31}\frac{1}{(1+\tau)^2}\prod_{i=1}^{12}\prod_{j=30}^{80}\left\{\binom{n_{ij}}{y_{ij}}\theta_{ij}^{y_{ij}}(1-\theta_{ij})^{n_{ij}-y_{ij}}\right\} \\ & \prod_{i=1}^{12}\left\{\prod_{j=30}^{k}\frac{\theta_{ij}^{\mu_1\tau-1}(1-\theta_{ij})^{(1-\mu_1)\tau-1}}{B(\mu_1\tau, (1-\mu_1)\tau)}\right. \\ & \left.\prod_{j=k+1}^{80}\frac{\theta_{ij}^{\mu_2\tau-1}(1-\theta_{ij})^{(1-\mu_2)\tau-1}}{B(\mu_2\tau, (1-\mu_2)\tau)}\right\}, \end{aligned} \tag{7}$$

and

$$p(\underset{\sim}{y}|k) = \frac{1}{31}\prod_{i=1}^{12}\prod_{j=30}^{80}\binom{n_{ij}}{y_{ij}}\int_0^\infty\int_0^1\int_0^1\frac{1}{(1+\tau)^2} \tag{8}$$

$$\prod_{i=1}^{12}\left\{\prod_{j=30}^{k}\frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)}\right. \tag{9}$$

$$\left.\prod_{j=k+1}^{80}\frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)}\right\}d\mu_1 d\mu_2 d\tau \tag{10}$$

Following the same procedure we presented in Chapter 2, we can use the Monto Carlo integration and results from the Gibbs sampler to compute the posterior distribution for k using all the data.

The posterior distribution for k is

$$p(k|\underset{\sim}{y}) = \frac{P(k=a_r)P(y \mid k=a_r)}{\sum_{s=40}^{70}P(k=a_s)P(y \mid k=a_s)} = \frac{P(y \mid k=a_r)}{\sum_{a_s=40}^{70}P(y \mid k=a_s)}, \quad a_r = 40,\ldots,70. \tag{11}$$

Once we obtain the distribution of k, we can make inferences about $\underset{\sim}{\theta}$. We can first draw k from $P(k|\underset{\sim}{y})$, and for each given k, draw μ_1, μ_2, τ from $P(\mu_1, \mu_2, \tau|k, \underset{\sim}{y})$. Then $\underset{\sim}{\theta}$ can be drawn from the following distributions,

$$\begin{aligned}
\theta_{ij} \mid \mu_1, \tau, k, \underset{\sim}{y} &\overset{ind}{\sim} \text{Beta}\,(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau) \qquad j = 30,\ldots,k \\
\theta_{ij} \mid \mu_2, \tau, k, \underset{\sim}{y} &\overset{ind}{\sim} \text{Beta}\,(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau) \qquad j = k+1,\ldots,80, i = 1,\ldots,12.
\end{aligned} \tag{12}$$

As presented in Table 4.1, we can observe that the change point for the unweighted data occurs around age 48. For the weighted data, the change point is about age 57. Figures 4.1 to 4.4 are plots for posterior mean and the 95% credible bands for θ_{ij}'s and the observed points.

Figure 4.1: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where only “Limited in major activity ” is considered as positive ALS.

k	case 1 unwt	case 2 unwt	case 1 logwt	case 2 logwt
40	0.00041	0.00005	0	0
41	0.00027	0.00005	0	0
42	0.00154	0.00014	0	0
43	0.00669	0.00127	0	0
44	0.00281	0.0004	0	0
45	0.06627	0.1383	0	0
46	0.10483	0.28211	0	0
47	0.13151	0.21517	0	0
48	**0.34779**	**0.32333**	0.00001	0
49	0.11675	0.02941	0.00009	0.00001
50	0.19466	0.00892	0.00074	0.00008
51	0.02574	0.00083	0.00363	0.0004
52	0.00069	0.00002	0.01208	0.00269
53	0.00004	0	0.04859	0.01273
54	0.00001	0	0.08612	0.0458
55	0	0	0.09549	0.03993
56	0	0	0.21706	0.15509
57	0	0	**0.23726**	**0.54696**
58	0	0	0.09808	0.10208
59	0	0	0.18147	0.09097
60	0	0	0.01651	0.00261
61	0	0	0.00276	0.00062
62	0	0	0.00011	0.00002
63	0	0	0.00001	0
64	0	0	0	0
65	0	0	0	0
66	0	0	0	0
67	0	0	0	0
68	0	0	0	0
69	0	0	0	0
70	0	0	0	0

Table 4.1: Distributions for the change point k for case 1 and case 2, using both unweighted and logistic weighted data, and assuming the same k for different domains. Case 1: Only "Limited in major activity" is considered as positive ALS. Case 2: Both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

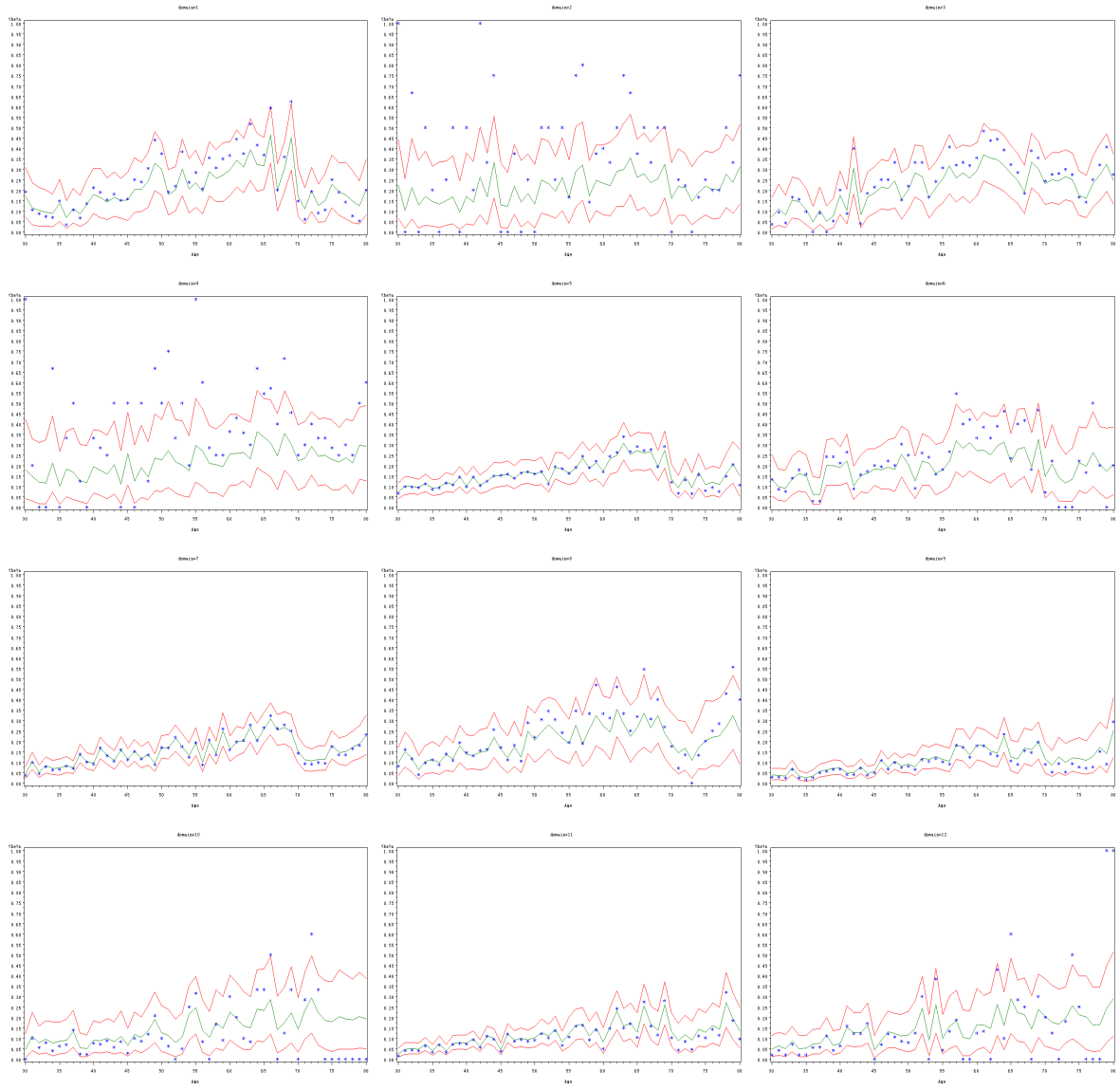

Figure 4.2: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where both “Limited in major activity” and “Limited in kind/amount of major activity ” are considered as positive ALS.

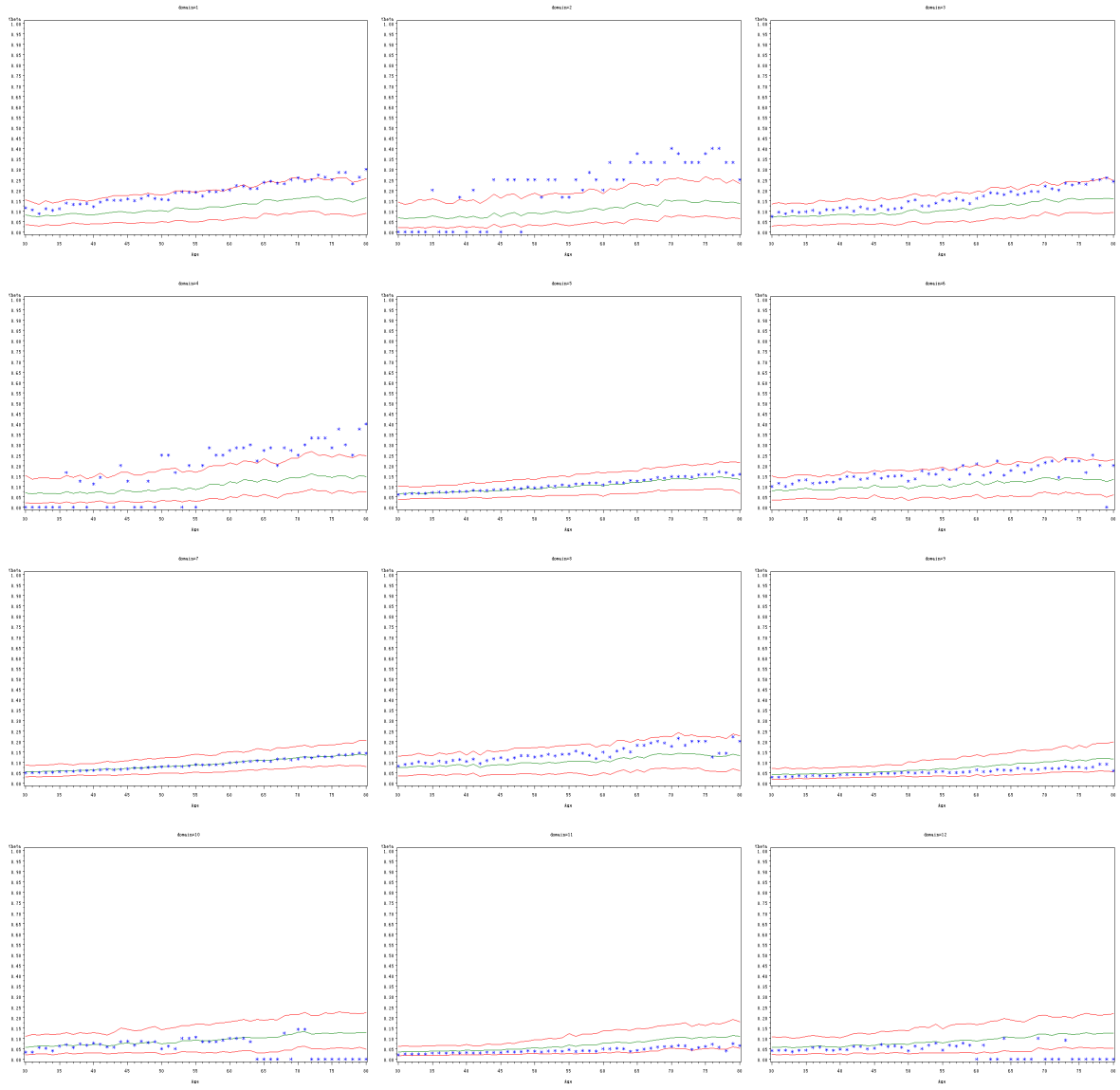

Figure 4.3: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

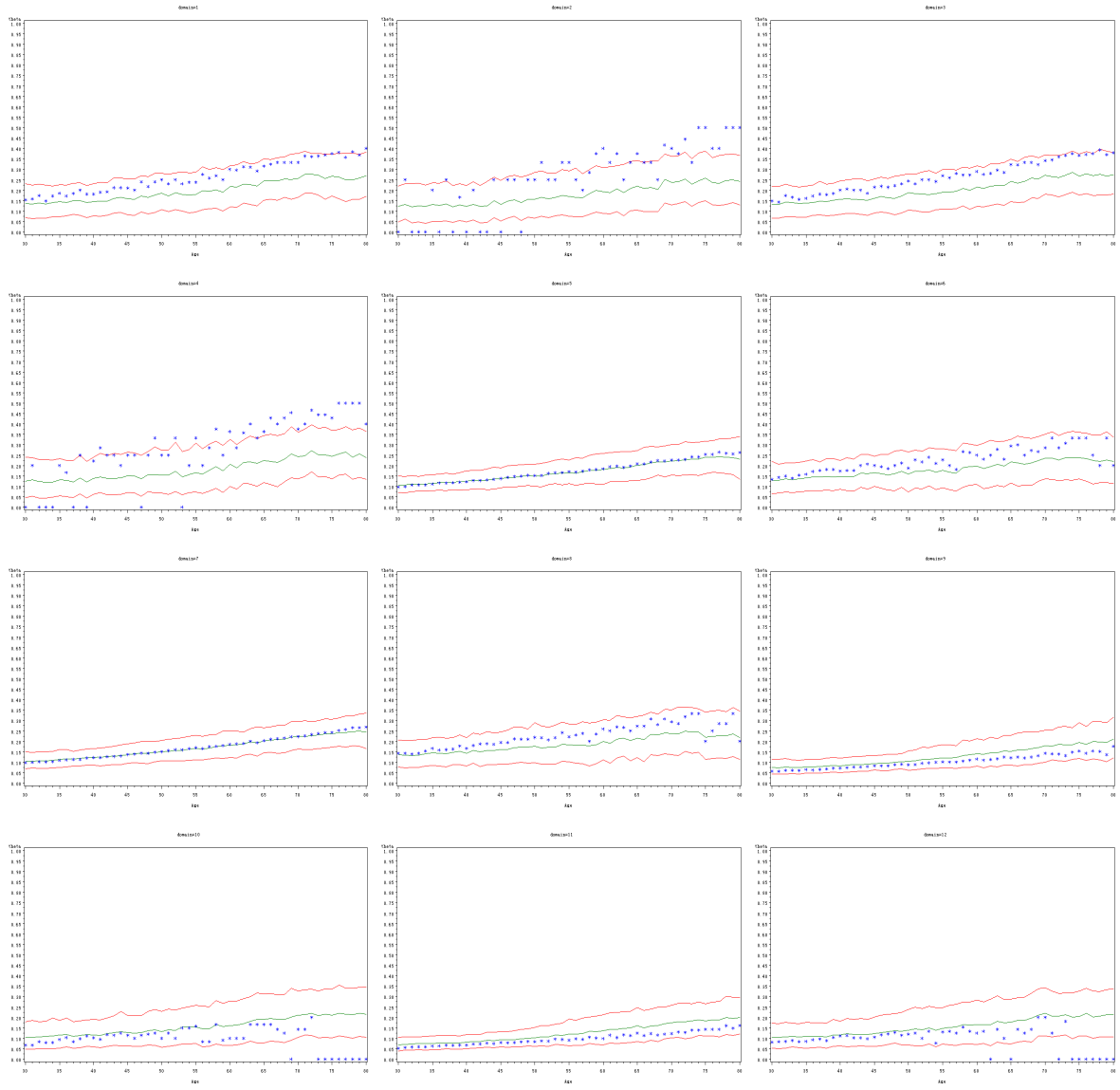

Figure 4.4: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

4.2 Different change points for the domains

In the second step, we assume that each domain has a different change point.So we will have a single model containing all the domains with different change points. Let $\underset{\sim}{k} = (k_1, k_2, \ldots, k_{12})$ denote the vector of change points, where $k_i, i = 1, \ldots, 12$ denote the change point for the i^{th}

domain and $\underset{\sim}{k}_{(i)}$ denotes the vector of change points for all the domains except the i^{th} domain. In addition to part 1, we are assuming that the k_i are identical and independently distributed.

Then the joint posterior distribution for $\underset{\sim}{\theta}$, μ_1, μ_2, and τ can be written as

$$p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau | \underset{\sim}{y}, \underset{\sim}{k}) \propto \frac{1}{(1+\tau)^2}\left(\frac{1}{31}\right)^{12}\prod_{i=1}^{12}\left\{\prod_{j=30}^{80}\binom{n_{ij}}{y_{ij}}\theta_{ij}^{y_{ij}}(1-\theta_{ij})^{n_{ij}-y_{ij}}\right.$$
$$\left.\prod_{j=30}^{k_i}\left[\frac{\theta_{ij}^{\mu_1\tau-1}(1-\theta_{ij})^{(1-\mu_1)\tau-1}}{B(\mu_1\tau,(1-\mu_1)\tau)}\right]\prod_{j=k_i+1}^{80}\left[\frac{\theta_{ij}^{\mu_2\tau-1}(1-\theta_{ij})^{(1-\mu_2)\tau-1}}{B(\mu_2\tau,(1-\mu_2)\tau)}\right]\right\} \quad (1)$$

and integrate over $\underset{\sim}{\theta}$ we can get

$$p(\mu_1, \mu_2, \tau | \underset{\sim}{y}, \underset{\sim}{k}) \propto \frac{1}{(1+\tau)^2}\left(\frac{1}{31}\right)^{12}\prod_{i=1}^{12}\left\{\prod_{j=30}^{80}\binom{n_{ij}}{y_{ij}}\right.$$
$$\prod_{j=30}^{k_i}\left[\frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)}\right]$$
$$\left.\prod_{j=k_i+1}^{80}\left[\frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)}\right]\right\}, \quad (2)$$

Now we want to compute the posterior distribution for $\underset{\sim}{k}$. However, this is a difficult problem. The procedure described in Chapter 2 is impractical here, because the computation is enormous. This is true because we have to, a) run 31^{12} Gibbs samplers, b) compute 31^{12} marginal likelihoods, and c) as we see in Chapter 3 the reversible jump sampler does not work.

Clearly, the computation is prohibitively expensive. To solve this problem, we can draw $p(\underset{\sim}{k}|\underset{\sim}{y})$ using a Gibbs sampler by drawing from $p(k_i|\underset{\sim}{k}_{(i)}, \underset{\sim}{y})$, where $\underset{\sim}{k}_{(i)}$ is the vector of the change points for each domain except the i^{th} domain.

$$p(k_i|\underset{\sim}{y}, \underset{\sim}{k}_{(i)}) \propto \int_0^\infty \int_0^1 \int_0^1 \prod_{i=1}^{12} \left\{ \prod_{j=30}^{k_i} \frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \right.$$
$$\left. \prod_{j=k_i+1}^{80} \frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \right\}$$
$$\frac{p(\mu_1,\mu_2,\tau|k)}{(1+\tau)^2 p(\mu_1,\mu_2,\tau|k)} d\mu_1 d\mu_2 d\tau \tag{3}$$

This can also be written as

$$p(k_i|\underset{\sim}{y}, \underset{\sim}{k}_{(i)}) \propto \int_0^\infty \int_0^1 \int_0^1 \prod_{j=30}^{k_i} \frac{B(y_{ij}+\mu_1\tau, n_{ij}-y_{ij}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)}$$
$$\prod_{j=k_i+1}^{80} \frac{B(y_{ij}+\mu_2\tau, n_{ij}-y_{ij}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)}$$
$$\prod_{l\neq i} \left\{ \prod_{j=30}^{k_i} \frac{B(y_{lj}+\mu_1\tau, n_{lj}-y_{lj}+(1-\mu_1)\tau)}{B(\mu_1\tau,(1-\mu_1)\tau)} \right.$$
$$\left. \prod_{j=k_l+1}^{80} \frac{B(y_{lj}+\mu_2\tau, n_{lj}-y_{lj}+(1-\mu_2)\tau)}{B(\mu_2\tau,(1-\mu_2)\tau)} \right\}$$
$$\frac{p(\mu_1,\mu_2,\tau|k)}{(1+\tau)^2 p(\mu_1,\mu_2,\tau|k)} d\mu_1 d\mu_2 d\tau \tag{4}$$

Frist, set $\underset{\sim}{k}_{(1)}$ to the starting value which we obtained from the first part. Using the importance function that we obtained in part 1, we can compute $p(k_1|\underset{\sim}{k}_{(1)}, \underset{\sim}{y})$ by Monte Carlo integration and Gibbs sampler for each $k_1 = 40, \ldots, 70$. Once we get this conditional posterior 'density' of $k_1|\underset{\sim}{k}_{(1)}, \underset{\sim}{y}$, we can draw a random value from it and fix the k_1 in $\underset{\sim}{k}_{(2)}$ at this value. So in $\underset{\sim}{k}_{(2)}$, k_1 will be the sample drawn from $p(k_1|\underset{\sim}{k}_{(1)}, \underset{\sim}{y})$ and $k_3, \ldots, k_{12}$ will still be the starting value we obtained from part 1. Now for the fixed k_1 and $\underset{\sim}{k}_{(1)}$, perform the Gibbs sample on $p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau|k_1, \underset{\sim}{k}_{(1)})$ at each value of k_2. At this point, k_2 is the starting value as we obtained in part 1. Obtain $p(k_2|\underset{\sim}{k}_{(2)}, \underset{\sim}{y})$ as in the previous step, and draw a value for $k_2|\underset{\sim}{k}_{(2)}, \underset{\sim}{y}$. Now we have updated k_1, k_2. We continue the process to update $k_3, k_4, \ldots, k_{12}$ in the same manner. Repeat the entire process until we get a large sample $\underset{\sim}{k}^{(1)}, \ldots, \underset{\sim}{k}^{(M)}$. This is the posterior density of

$p(\underset{\sim}{k}, \underset{\sim}{y})$. Now we can construct 95% credible interval for each component $k_1, \ldots, k_{12}$.

The next step is to make inference about $\underset{\sim}{\theta}$'s. The posterior density of $p(\underset{\sim}{\theta}|\mu_1, \mu_2, \tau, \underset{\sim}{k}, \underset{\sim}{y})$ is straight forwad to obtain.

$$p(\underset{\sim}{\theta}, \mu_1, \mu_2, \tau, \underset{\sim}{k}|\underset{\sim}{y}) = p(\underset{\sim}{\theta}|\mu_1, \mu_2, \tau, \underset{\sim}{k}, \underset{\sim}{y})p(\mu_1, \mu_2, \tau|\underset{\sim}{k}, \underset{\sim}{y})p(\underset{\sim}{k}|\underset{\sim}{y}) \tag{5}$$

Note that $p(\mu_1, \mu_2, \tau|\underset{\sim}{k}, \underset{\sim}{y})$ is not in close form. We can draw $\underset{\sim}{\theta}, \mu_1, \mu_2, \tau|\underset{\sim}{k}, \underset{\sim}{y}$ in two steps.

First, use the Metropolis sampler to draw $\mu_1, \mu_2, \tau|\underset{\sim}{k}, \underset{\sim}{y}$ at each value of $\underset{\sim}{k}$, (i.e., $\underset{\sim}{k}^{(1)}, \ldots, \underset{\sim}{k}^{(M)}$). Assume Beta distributions for $\underset{\sim}{\mu}$, and Gamma distributions for τ as candicate generating densities, we have the candidate generating density $p_a(\mu_1, \mu_2, \tau|\underset{\sim}{k}, \underset{\sim}{y})$,

$$\mu_i \overset{iid}{\sim} \text{Beta}(\nu_i\phi_i, (1-\nu_i)\phi_i) \quad i = 1, 2 \tag{6}$$

$$\tau \sim \text{Gamma}(\alpha, \beta), \tag{7}$$

where,

$$\nu_i = \frac{1}{M}\sum_{h=1}^{M} \mu_i^{(h)} \tag{8}$$

$$\phi_i = \frac{\nu_i(1-\nu_i)}{\frac{1}{M-1}\sum_{h=1}^{M}(\mu_i^{(h)} - \frac{1}{M}\sum_{h=1}^{M}\mu_i^{(h)})^2} \tag{9}$$

$$\alpha = \beta\frac{1}{M}\sum_{h=1}^{M}\tau^{(h)} \tag{10}$$

$$\beta = \frac{\frac{1}{M}\sum_{h=1}^{M}\tau^{(h)}}{\frac{1}{M-1}\sum_{h=1}^{M}(\tau^{(h)} - \frac{1}{M}\sum_{h=1}^{M}\tau^{(h)})^2}, \tag{11}$$

and $(\mu_i^{(h)}, \tau^{(h)}), h = 1, \ldots, M, M \approx 1000$ are obtained from the Gibbs sampler.

Let Ω denote (μ_1, μ_2, τ), and draw samples Ω_0, Ω_1 from the above distributions. Then we compute the M-H sampler

$$\alpha(\Omega_0, \Omega_1) = \min\left\{1, \frac{\pi(\Omega_1)}{\pi(\Omega_0)}\right\}, \tag{12}$$

where

$$\pi(\Omega) = \frac{p(\mu_1, \mu_2, \tau | k, \underset{\sim}{y})}{p_a(\mu_1, \mu_2, \tau | \underset{\sim}{k}, \underset{\sim}{y})}. \tag{13}$$

Draw a random number from Uniform(0,1) distribution. If this number is smaller than or equal to $\alpha(\Omega_0, \Omega_1)$, then we take Ω_1, other wise we stay at Ω_0. Repeat this whole process and we can obtain samples of $\mu_1^{(h)}, \mu_2^{(h)}, \tau^{(h)} | \underset{\sim}{k}, \underset{\sim}{y}$, $h = 1, \ldots, M$.

After we have samples of $\mu_1^{(h)}, \mu_2^{(h)}, \tau^{(h)}$, and $k^{(h)}$, we can fill in the $\underset{\sim}{\theta}$'s from Beta distributions as we have done before.

$$\begin{aligned}
\theta_{ij} \mid \mu_1, \tau, k, \underset{\sim}{y} &\overset{ind}{\sim} \text{Beta}\left(y_{ij} + \mu_1\tau, n_{ij} - y_{ij} + (1-\mu_1)\tau\right) \qquad j = 30, \ldots, k \\
\theta_{ij} \mid \mu_2, \tau, k, \underset{\sim}{y} &\overset{ind}{\sim} \text{Beta}\left(y_{ij} + \mu_2\tau, n_{ij} - y_{ij} + (1-\mu_2)\tau\right) \qquad j = k+1, \ldots, 80, i = 1, \ldots, 12.
\end{aligned} \tag{14}$$

The posterior distributions for k are presented in Tables 4.2-4.5. For the unweighted data, the distributions for the change points center around age 40 for the first 4 domain, which corresponds to adults with low education level. The change points for domains 5-8 occur around age 40 to 45, which is later than the first 4 domains. These domains corresponds to adults with median education levels. The last four domains, which are adults with high education, have much later onsets around age 50 to 60. This shows that the onset of activity limitation is sensitive to adults' education levels, which is the same to the conclusion we have obtained in the previous chapters. For the logistic weighted data, the trend is the same. The first four domains (low education) have early onsets around age 40. Domains 5-8 (median education) have later onsets around 40 to 60. The onsets for the last four domains (high education) occur around age 70. The definition of positive ALS, whether to include "Limited in kind/amount of major activity limitation" does not make a large difference here. Comparing with the distributions for k obtained in Chapter 2, these distributions are more concentrated, and therefore, more accurate.

For each case, the plots for the mean and 95% credible bands together with observed data are presented in Figures 4.5-4.8. Most observed data points fall between the credible bands, which indicates that this is not a bad fit. And as before, the plots with logistic weighted data is more smooth than the plots using the unweighted data.

The apparent increase of change point with education needs to be explored further. We observe that many of the cell counts for black females for low and high education are zeros. So these results are suspectible. One might need to collapse over race for low and high education to see if these results prevail.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
40	**0.334**	**0.375**	0.25	0.133	**0.222**	**0.385**	0.055	**0.182**	0	0.022	0.003	0.001
41	0.21	0.288	**0.262**	0.085	0.154	0.175	0.071	0.174	0	0.024	0.001	0.001
42	0.156	0.126	0.112	0.086	0.179	0.156	0.066	0.176	0	0.014	0.011	0.002
43	0.102	0.067	0.112	**0.142**	0.105	0.122	0.078	0.102	0	0.012	0.007	0.006
44	0.068	0.027	0.131	**0.142**	0.091	0.07	0.081	0.05	0.001	0.011	0.008	0.007
45	0.054	0.022	0.055	0.067	0.075	0.035	0.08	0.046	0.004	0.027	0.015	0.007
46	0.04	0.03	0.031	0.058	0.067	0.019	0.046	0.046	0.006	0.029	0.009	0.022
47	0.018	0.006	0.018	0.083	0.025	0.014	0.073	0.044	0.011	0.028	0.023	0.012
48	0.009	0.021	0.015	0.071	0.034	0.013	0.074	0.051	0.019	0.02	0.021	0.023
49	0.002	0.012	0.004	0.037	0.015	0.007	**0.091**	0.061	0.038	0.029	0.039	0.042
50	0.001	0.014	0.004	0.046	0.012	0.001	0.095	0.033	0.032	0.022	0.042	**0.099**
51	0.001	0.008	0.002	0.019	0.004	0	0.069	0.019	0.038	0.047	0.06	0.082
52	0.003	0.002	0.001	0.009	0.005	0.001	0.031	0.011	0.046	0.061	0.055	0.049
53	0	0.001	0.002	0.006	0.006	0	0.017	0.003	0.043	0.065	0.043	0.07
54	0	0.001	0	0.009	0.004	0	0.019	0.001	0.074	0.047	0.063	0.041
55	0.002	0	0	0.004	0	0	0.016	0	0.082	0.034	**0.08**	0.064
56	0	0	0	0	0.001	0.001	0.014	0.001	**0.125**	0.041	0.072	0.038
57	0	0	0.001	0.001	0.001	0.001	0.01	0	0.07	0.073	0.043	0.036
58	0	0	0	0.002	0	0	0.002	0	0.05	0.054	0.057	0.047
59	0	0	0	0	0	0	0.004	0	0.082	0.046	0.047	0.064
60	0	0	0	0	0	0	0.005	0	0.041	0.024	0.066	0.058
61	0	0	0	0	0	0	0.001	0	0.045	0.033	0.044	0.043
62	0	0	0	0	0	0	0.002	0	0.034	0.048	0.037	0.04
63	0	0	0	0	0	0	0	0	0.035	0.035	0.029	0.019
64	0	0	0	0	0	0	0	0	0.014	0.034	0.02	0.037
65	0	0	0	0	0	0	0	0	0.024	0.04	0.034	0.019
66	0	0	0	0	0	0	0	0	0.029	0.013	0.034	0.018
67	0	0	0	0	0	0	0	0	0.022	0.014	0.019	0.011
68	0	0	0	0	0	0	0	0	0.009	0.011	0.01	0.017
69	0	0	0	0	0	0	0	0	0.011	0.019	0.006	0.016
70	0	0	0	0	0	0	0	0	0.015	0.023	0.002	0.009

Table 4.2: Distributions for the change point k for the 12 domains using the revised Bayesian hierachical model and unweighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
40	**0.235**	**0.42**	0.166	**0.303**	0.038	**0.231**	0.012	0.136	0	0	0	0
41	0.158	0.391	**0.331**	0.214	0.027	0.083	0.01	**0.183**	0	0	0	0
42	0.169	0.078	0.06	0.148	0.071	0.168	0.01	0.171	0	0	0	0
43	0.126	0.051	0.183	0.044	0.081	0.156	0.024	0.162	0	0.001	0	0
44	0.116	0.004	0.127	0.117	0.076	0.133	0.016	0.059	0	0.001	0	0
45	0.105	0.01	0.07	0.034	0.079	0.092	0.024	0.046	0	0.002	0	0
46	0.057	0.019	0.029	0.05	0.077	0.062	0.021	0.069	0	0	0.001	0
47	0.02	0.001	0.021	0.037	**0.093**	0.03	0.069	0.038	0	0.002	0.002	0.001
48	0.008	0.009	0.002	0.032	0.069	0.02	0.055	0.09	0	0.003	0.001	0.002
49	0.003	0.003	0.005	0.014	0.059	0.008	**0.168**	0.024	0.001	0.003	0.003	0.006
50	0.001	0.011	0.004	0.005	0.063	0.005	0.116	0.018	0.001	0.002	0.002	0.009
51	0	0.002	0.001	0.002	0.043	0.008	0.1	0.003	0.007	0.014	0.001	0.005
52	0.002	0.001	0	0	0.087	0.001	0.045	0.001	0.012	0.037	0.001	0.004
53	0	0	0	0	0.053	0	0.026	0	0.015	**0.109**	0.003	0.012
54	0	0	0.001	0	0.032	0.001	0.054	0	0.012	0.054	0.011	0.002
55	0	0	0	0	0.026	0	0.029	0	0.036	0.016	0.019	0.01
56	0	0	0	0	0.012	0.001	0.105	0	0.086	0.032	0.012	0.013
57	0	0	0	0	0.004	0.001	0.035	0	0.044	0.094	0.015	0.007
58	0	0	0	0	0.008	0	0.045	0	0.036	0.08	0.029	0.025
59	0	0	0	0	0	0	0.012	0	0.056	0.1	0.039	0.115
60	0	0	0	0	0.002	0	0.011	0	0.04	0.058	**0.178**	0.124
61	0	0	0	0	0	0	0.007	0	0.031	0.046	0.158	0.13
62	0	0	0	0	0	0	0.006	0	0.025	0.072	0.078	**0.216**
63	0	0	0	0	0	0	0	0	0.041	0.088	0.072	0.082
64	0	0	0	0	0	0	0	0	0.013	0.052	0.063	0.111
65	0	0	0	0	0	0	0	0	0.037	0.034	0.146	0.037
66	0	0	0	0	0	0	0	0	0.093	0.013	0.037	0.016
67	0	0	0	0	0	0	0	0	0.09	0.019	0.027	0.019
68	0	0	0	0	0	0	0	0	**0.107**	0.024	0.062	0.023
69	0	0	0	0	0	0	0	0	0.054	0.008	0.018	0.019
70	0	0	0	0	0	0	0	0	0.163	0.036	0.022	0.012

Table 4.3: Distributions for the change point k for the 12 domains using the revised Bayesian hierachical model and unweighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
40	**0.579**	0.17	**0.252**	0.067	0	**0.644**	0	**0.288**	0	0	0	0
41	0.269	0.131	0.183	0.064	0	0.222	0	0.174	0	0	0	0
42	0.107	0.178	0.164	0.07	0.002	0.082	0	0.195	0	0	0	0
43	0.026	**0.223**	0.112	0.083	0.005	0.029	0	0.147	0	0.002	0	0
44	0.008	0.086	0.084	0.063	0.007	0.02	0	0.079	0	0	0	0
45	0.004	0.084	0.072	0.066	0.011	0.001	0	0.048	0	0	0	0
46	0.002	0.054	0.032	0.084	0.032	0	0	0.029	0	0	0	0
47	0.003	0.014	0.037	**0.091**	0.05	0	0	0.021	0	0	0	0
48	0	0.025	0.026	0.079	0.071	0	0	0.014	0	0	0	0
49	0.001	0.015	0.025	0.086	0.092	0.002	0	0.003	0	0	0	0
50	0.001	0.008	0.007	0.057	0.11	0	0	0.001	0	0.004	0	0
51	0	0.006	0.002	0.048	0.16	0	0	0	0	0.006	0	0
52	0	0.004	0.001	0.031	0.12	0	0.001	0.001	0	0.012	0	0
53	0	0.001	0.002	0.026	**0.125**	0	0.003	0	0	0.016	0	0
54	0	0.001	0	0.032	0.077	0	0.004	0	0	0.012	0	0
55	0	0	0.001	0.028	0.068	0	0.01	0	0	0.005	0	0
56	0	0	0	0.018	0.033	0	0.025	0	0	0.009	0	0.001
57	0	0	0	0.005	0.02	0	0.049	0	0	0.011	0	0
58	0	0	0	0.002	0.005	0	0.102	0	0	0.011	0	0.001
59	0	0	0	0	0.009	0	0.133	0	0	0.01	0	0.002
60	0	0	0	0	0.001	0	0.132	0	0	0.019	0	0.007
61	0	0	0	0	0.002	0	**0.147**	0	0	0.017	0	0.008
62	0	0	0	0	0	0	0.132	0	0	0.017	0	0.01
63	0	0	0	0	0	0	0.119	0	0	0.018	0	0.019
64	0	0	0	0	0	0	0.057	0	0	0.039	0	0.012
65	0	0	0	0	0	0	0.032	0	0	0.057	0	0.036
66	0	0	0	0	0	0	0.034	0	0.008	0.102	0.003	0.059
67	0	0	0	0	0	0	0.01	0	0.02	0.147	0.015	0.126
68	0	0	0	0	0	0	0.006	0	0.083	0.141	0.044	0.215
69	0	0	0	0	0	0	0.004	0	0.223	**0.179**	0.225	0.195
70	0	0	0	0	0	0	0	0	**0.666**	0.166	**0.713**	**0.309**

Table 4.4: Distributions for the change point k for the 12 domains using the revised Bayesian hierachical model and logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

Age	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12
40	**0.432**	0.076	**0.566**	**0.318**	0	**0.357**	0	**0.426**	0	0	0	0
41	0.306	0.051	0.212	0.182	0	0.27	0	0.262	0	0	0	0
42	0.163	0.11	0.103	0.119	0	0.216	0	0.152	0	0	0	0
43	0.06	**0.183**	0.055	0.076	0	0.09	0	0.084	0	0	0	0
44	0.015	0.1	0.036	0.083	0	0.036	0	0.038	0	0	0	0
45	0.01	0.129	0.017	0.044	0.001	0.011	0	0.022	0	0	0	0
46	0.007	0.105	0.006	0.038	0.004	0.011	0	0.013	0	0	0	0
47	0.002	0.061	0.002	0.048	0.007	0.006	0.001	0.002	0	0	0	0
48	0.003	0.068	0.002	0.034	0.013	0	0.005	0.001	0	0	0	0
49	0.001	0.05	0.001	0.019	0.031	0.001	0.013	0	0	0	0	0
50	0.001	0.028	0	0.01	0.065	0	0.039	0	0	0.002	0	0
51	0	0.018	0	0.008	0.108	0	0.089	0	0	0	0	0
52	0	0.01	0	0.003	0.14	0.002	0.109	0	0	0	0	0
53	0	0.007	0	0.008	**0.153**	0	0.152	0	0	0	0	0
54	0	0.002	0	0.006	0.128	0	0.137	0	0	0	0	0
55	0	0.001	0	0.001	0.105	0	0.14	0	0	0	0	0.001
56	0	0.001	0	0.003	0.108	0	**0.161**	0	0	0.003	0	0.001
57	0	0	0	0	0.068	0	0.076	0	0	0.01	0	0.006
58	0	0	0	0	0.037	0	0.045	0	0	0.007	0	0.005
59	0	0	0	0	0.016	0	0.02	0	0	0.015	0	0.006
60	0	0	0	0	0.013	0	0.009	0	0	0.022	0	0.007
61	0	0	0	0	0.002	0	0.004	0	0	0.029	0	0.01
62	0	0	0	0	0.001	0	0	0	0	0.065	0	0.022
63	0	0	0	0	0	0	0	0	0	0.076	0	0.032
64	0	0	0	0	0	0	0	0	0	0.082	0	0.046
65	0	0	0	0	0	0	0	0	0	0.072	0	0.128
66	0	0	0	0	0	0	0	0	0.002	0.07	0.003	0.132
67	0	0	0	0	0	0	0	0	0.024	0.066	0.005	**0.171**
68	0	0	0	0	0	0	0	0	0.091	0.104	0.042	**0.171**
69	0	0	0	0	0	0	0	0	0.255	0.171	0.204	0.132
70	0	0	0	0	0	0	0	0	**0.628**	**0.206**	**0.746**	0.13

Table 4.5: Distributions for the change point k for the 12 domains using the revised Bayesian hierachical model and logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

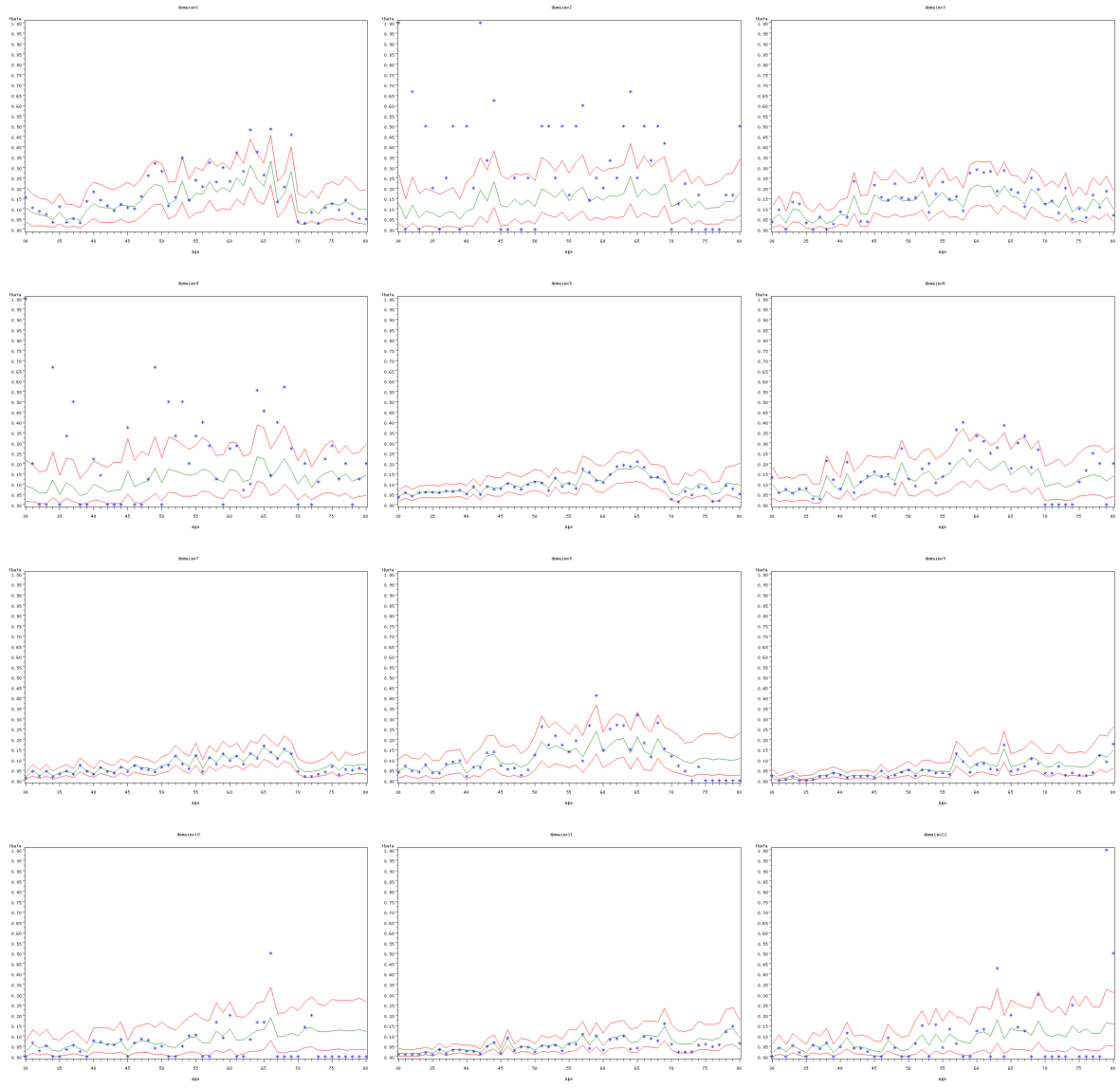

Figure 4.5: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where only “Limited in major activity ” is considered as positive ALS.

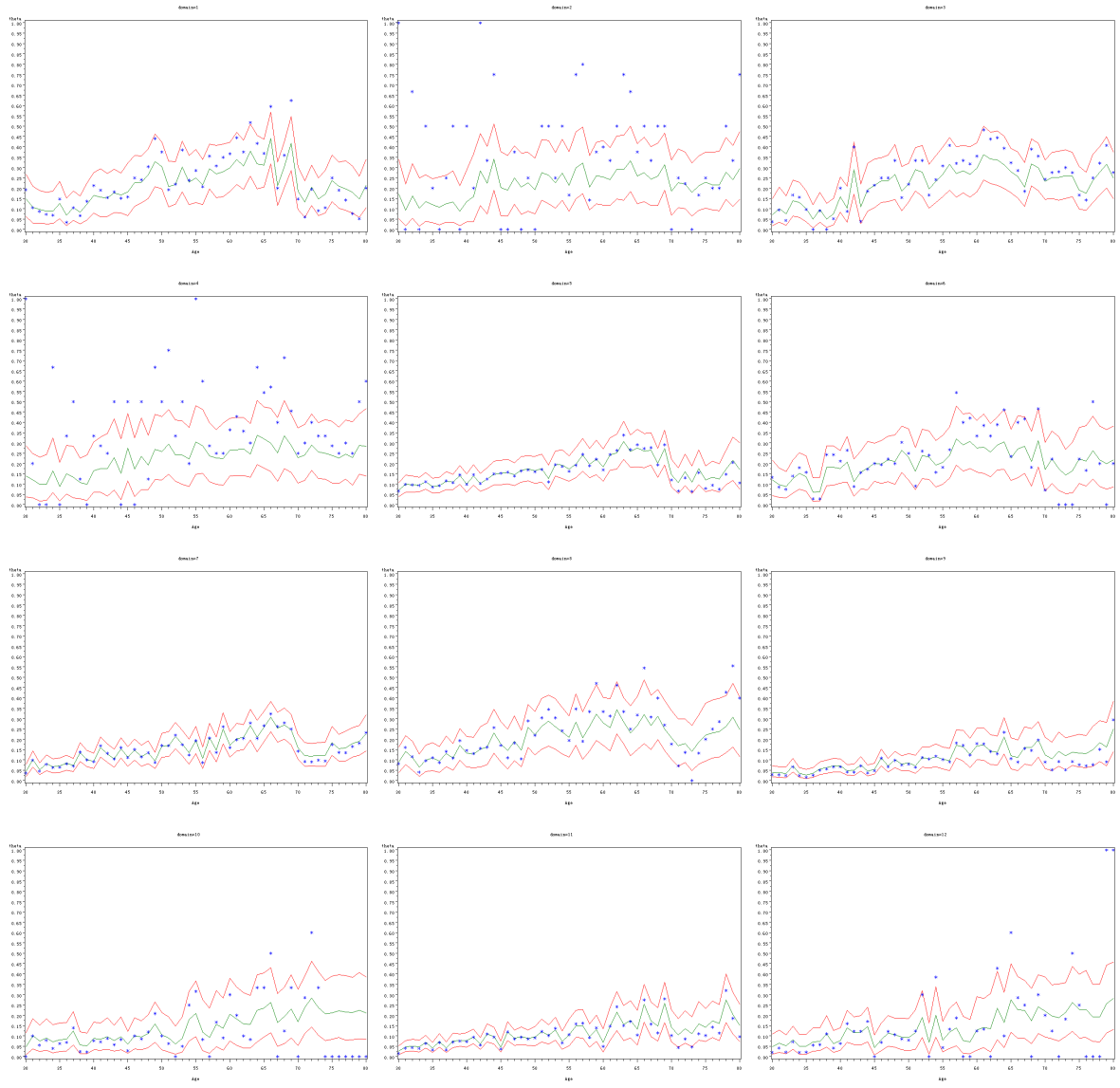

Figure 4.6: Plots for posterior mean and 95% credible bands with observed data for each domain using unweighted data, where both “Limited in major activity” and “Limited in kind/amount of major activity ” are considered as positive ALS.

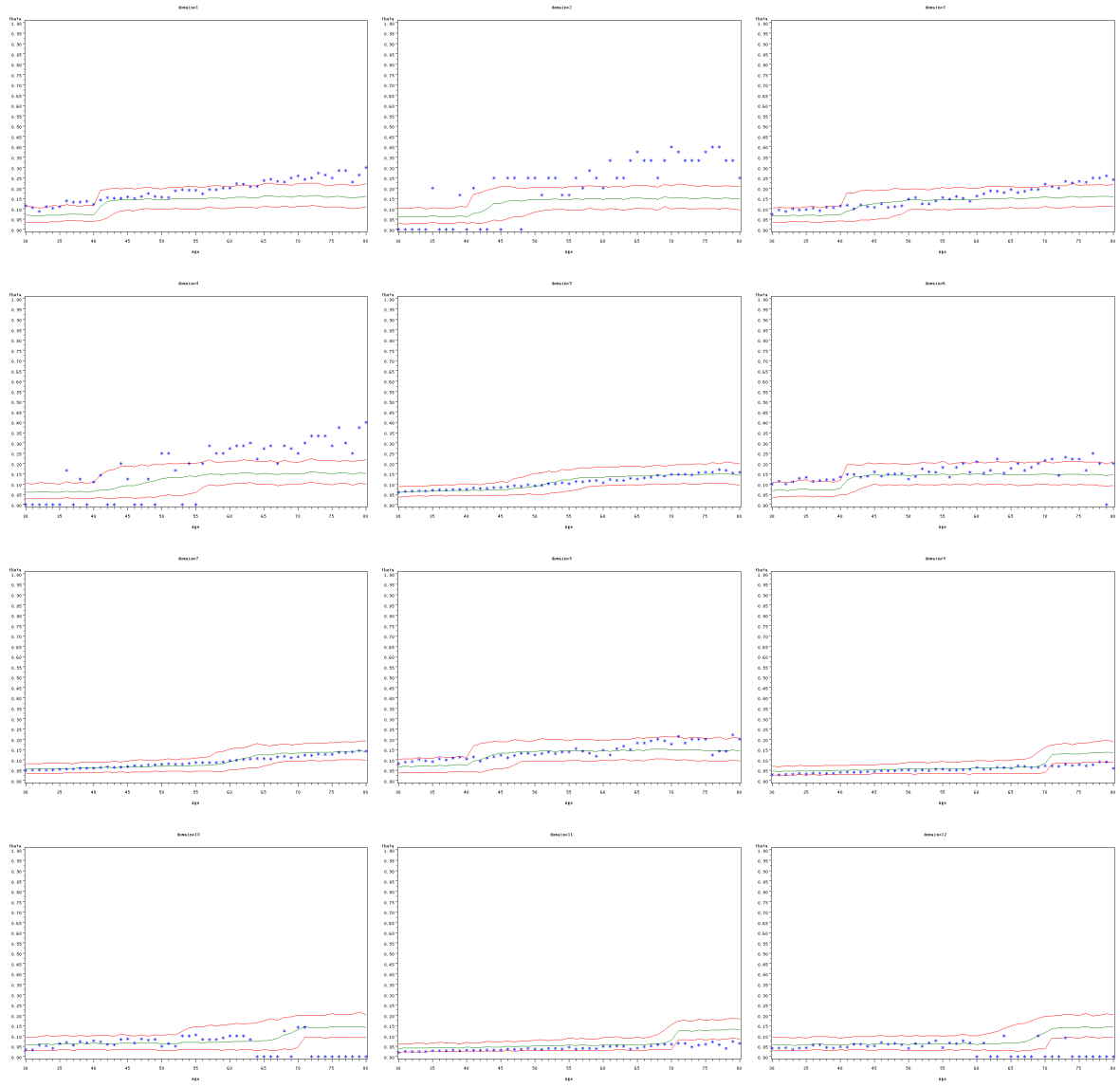

Figure 4.7: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

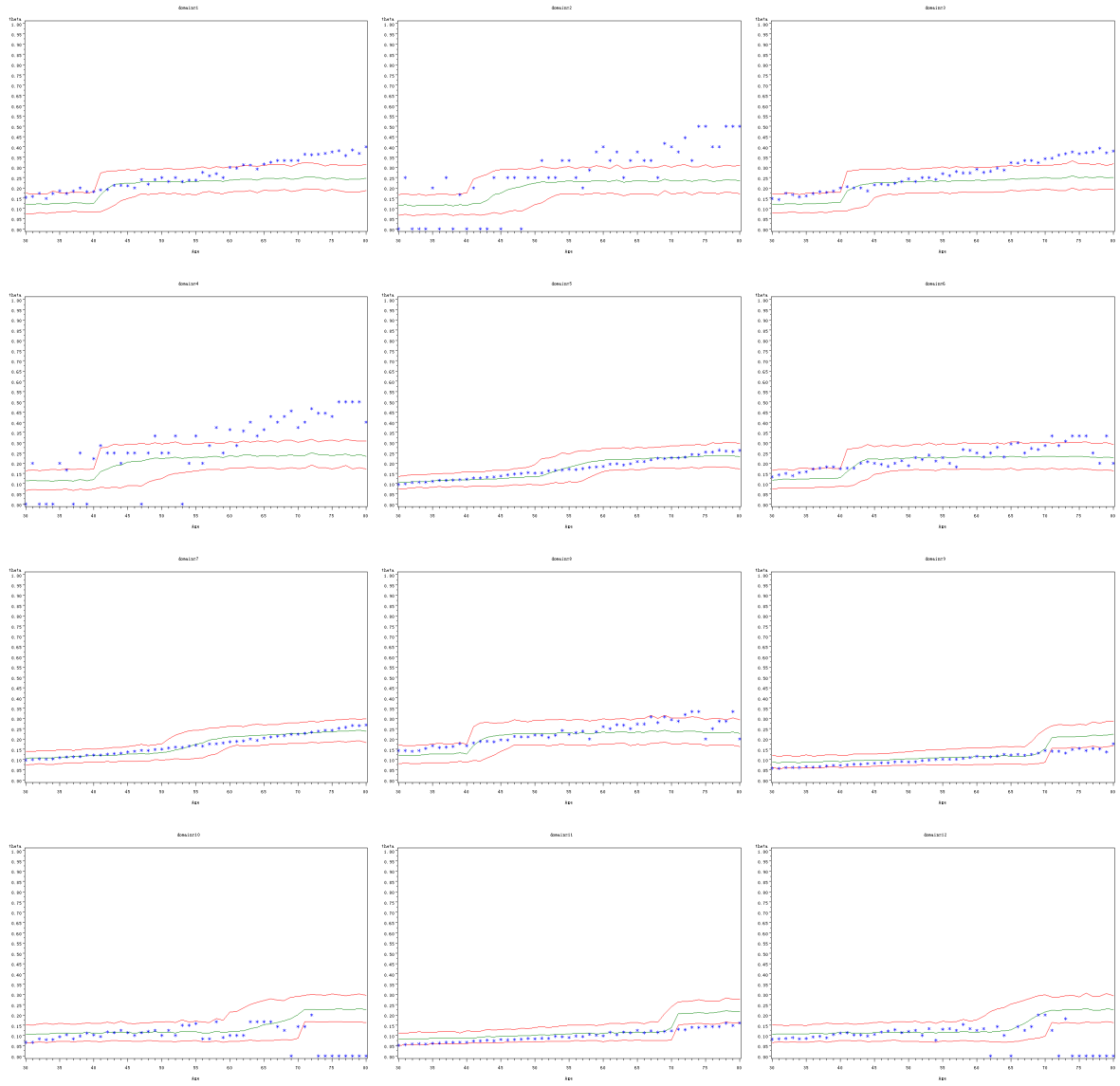

Figure 4.8: Plots for posterior mean and 95% credible bands with observed data for each domain using logistic weighted data, where both “Limited in major activity” and “Limited in kind/amount of major activity” are considered as positive ALS.

Chapter 5

Diagnostics of the Models

We have developed three models in the previous chapters. The reversible jump model has already been proved to be deficient in Chapter 3. To compare the models in Chapter 2 and Chapter 4, we use the diagnostic method called conditional predictive ordinate (CPO).

5.1 Cross valitaton deleted residuals

We first look at the cross validation deleted residuals, which are defined as

$$\text{cross validation deleted residual} = \frac{E(y_{ij}|\underset{\sim}{y}_{(ij)}) - y_{ij}}{\sqrt{Var(y_{ij}|\underset{\sim}{y}_{(ij)})}} \tag{1}$$

where $\underset{\sim}{y}_{(ij)}$ is the vector of all the y's excluding the ij^{th} observation. We plot the cross validation deleted residuals against the observed probabilities and examine the outliers. The plots for the two models and different cases are presented in Figures 5.1 and 5.2. We notice that the models using the logistic weighted data fit better than the models using the unweighted data because the deleted residuals are more close to 0. The two vertical streaks in each of the four graphs in Figure 5.2 is due to many close predicted values. When we blou them up, these steaks disappear.

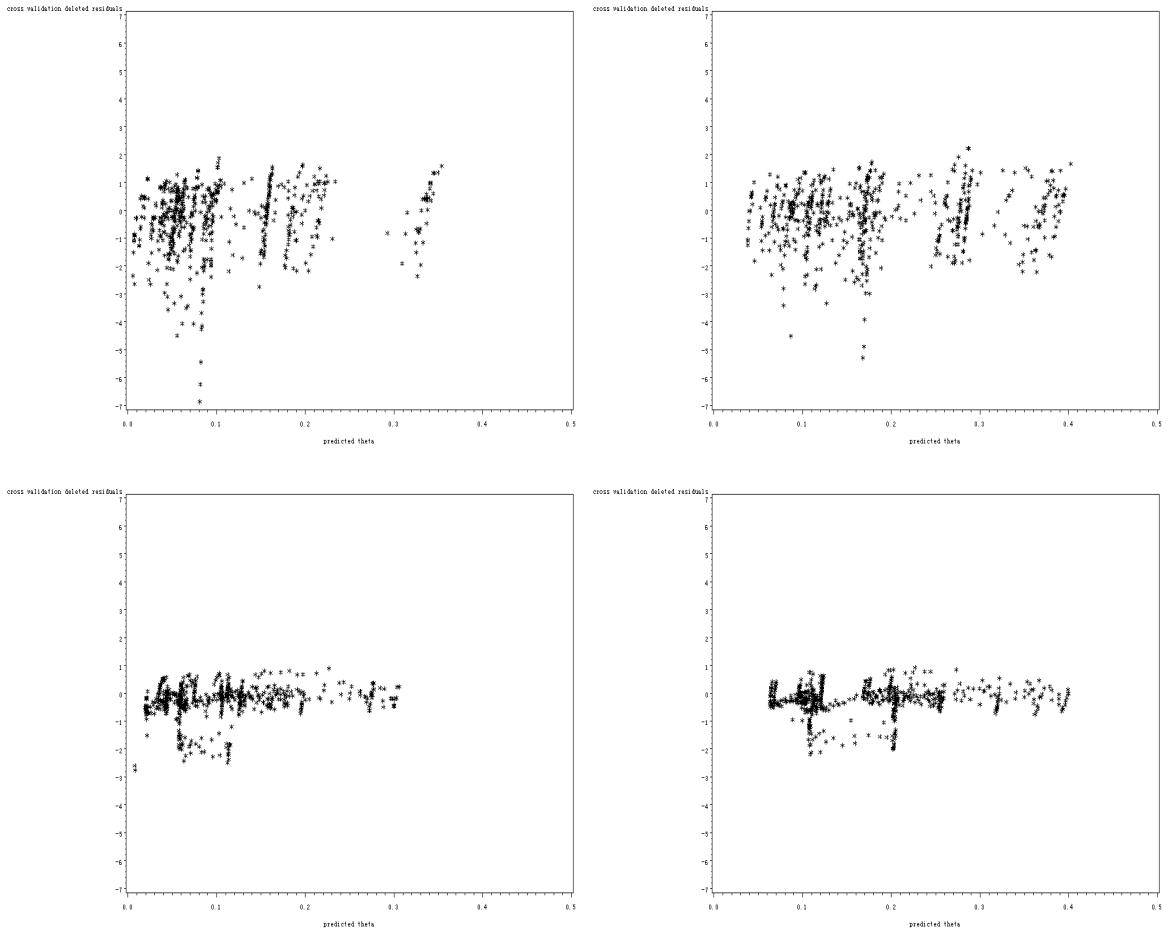

Figure 5.1: Plots of cross validation deleted residuals against the predicted θ's for the model in Chapter 2. Top panel: unweighted data; bottom panel: logistic weighted $\tilde{\text{data}}$.

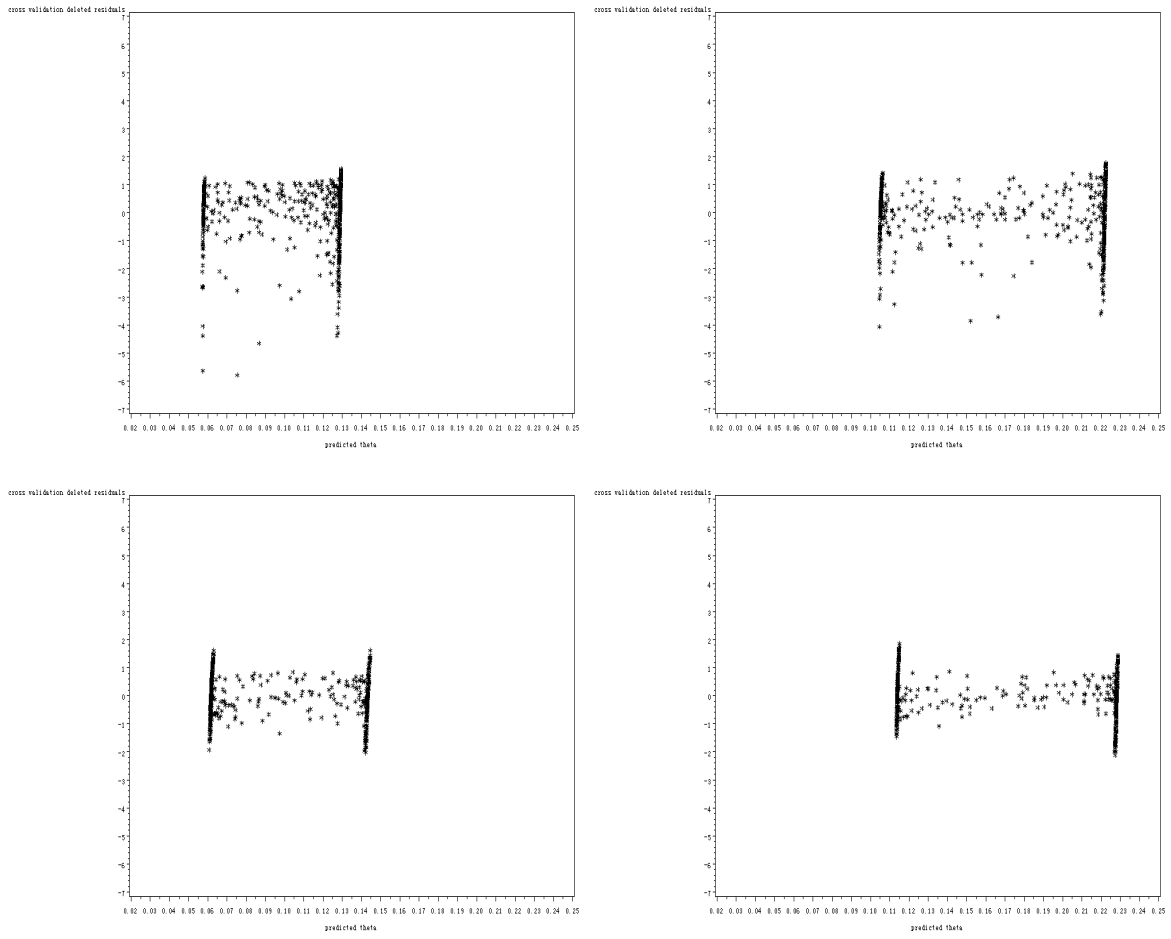

Figure 5.2: Plots of cross validation deleted residuals against the predicted θ's for the model in Chapter 4. Top panel: unweighted data; bottom panel: logistic weighted data.

5.2 Conditional predictive ordinates

The other criterion we look at is the conditional predictive ordinate (CPO). Let $p(y_{ij}|y_{(ij)})$, also known as the conditional predictive density, denote the probability of y_{ij} conditional on $y_{(ij)}$. If y_{ij} is a single outlier, the probability to predice y_{ij} given the rest of the sample is very low. Figure 5.3 and 5.4 are plots of the CPO for the two models. The plots indicate the second Bayesian hierarchical model is better than the first model because there are fewer outliers at extremely low values. Also, the models using the logistic data are better than those using the unweighted data. The averages of the logarithm CPO for each model are presented in Table

5.1. It appears that the first model is slightly better than the second model. However, the first model does not take care of the heterogeneity among the domains. Also, the models using the logistic weighted data fit much better than those using the unweighted data.

Model	Unweighted	Unweighted	Weighted	Weighted
	Case I	Case II	Case I	Case II
Model 1	-2.17	-2.42	-1.84	-2.01
Model 2	-2.24	-2.53	-1.85	-2.06

Table 5.1: Averages of CPO's for each model.

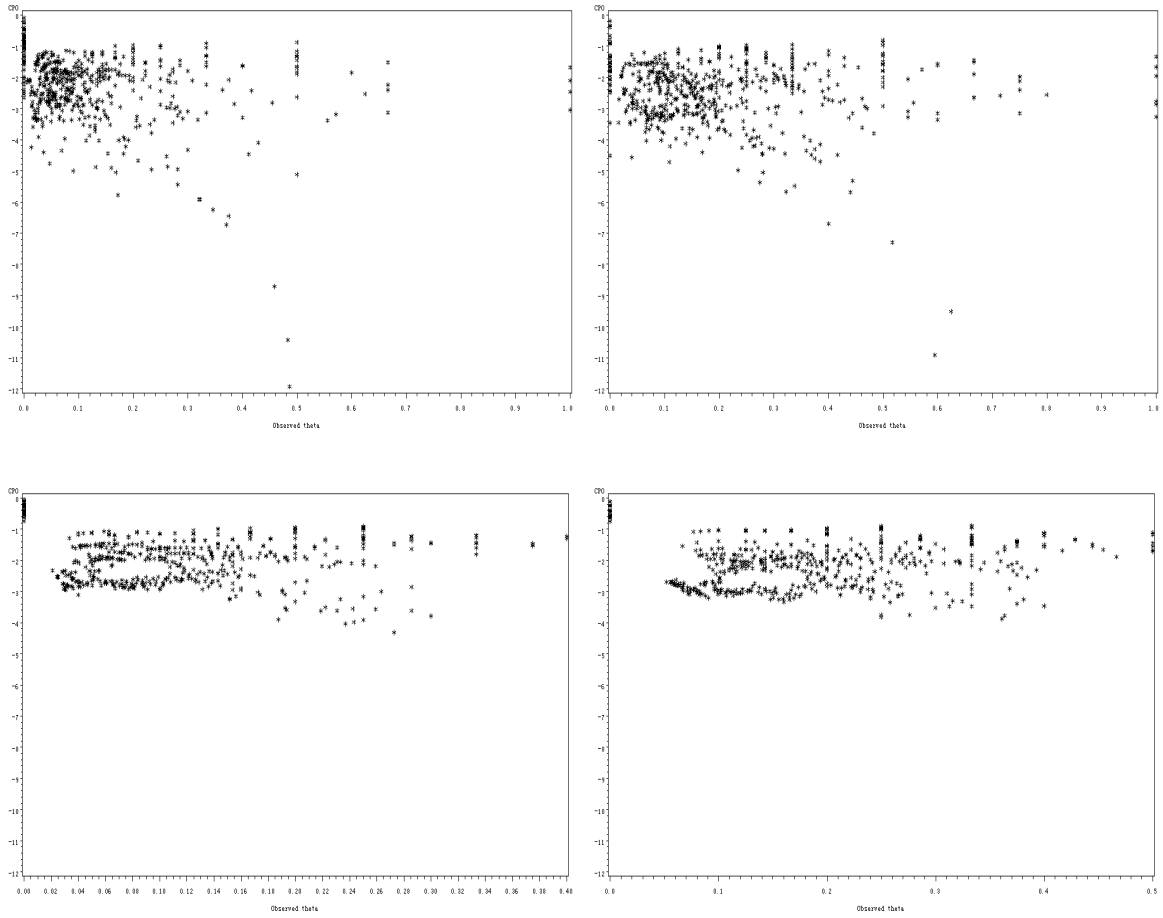

Figure 5.3: Plots of cross validation deleted residuals against the observed θ's for the model in Chapter 2. Top panel: unweighted data; bottom panel: logistic weighted data.

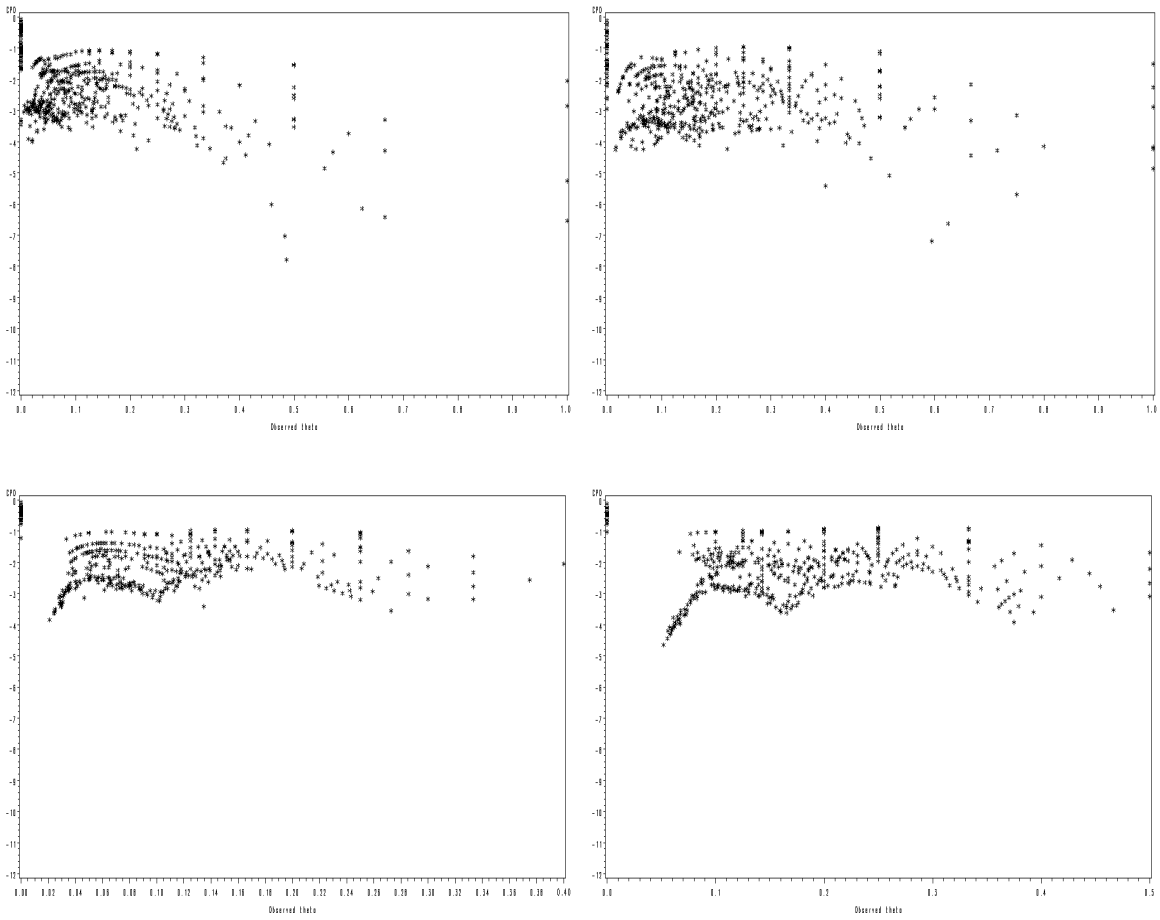

Figure 5.4: Plots of cross validation deleted residuals against the observed θ's for the model in Chapter 4. Top panel: unweighted data; bottom panel: logistic weighted data.

5.3 Collapsing

The results in the previous chapters indicate that the onset of ALS occurs around age 40 for adults with lower education, and occurs around age 70 for adults with higher education. There is a huge difference between these two groups and they are both near the border lines. Therefore, this might not be a very accurate reflection of the real onset of ALS. After a further examination of our data, we found that for some domains (e.g., 2, 4, 10, 12) the data are very sparse and there are even zero counts for population size. Therefore, the results of the relationship between change point and education level is suspect. To further explore this, we

collapse over race and sex to obtain the results based on 3 domains by education (low, median, high). The posterior distributions for the change point k using the Bayesian model pooling over domains and collapsed over race and sex are presented in Tables 7.2-7.5. We notice that the onsets for the unweighted data are early, around ages 40 to 50. The models using the logistic weighted data produce later onsets, which are around age 50 to 60. The difference among education levels is not large and we do not have the previous conclusion about the relationship between onset and education.

5.4 Conclusion

Based on the previous analysis, we can conclude that although the models for individuals domains perform slightly better than the single model assuming different change points for different domains, they do not take care of the herterogeneity among the domains. So we still prefer the single model assuming different change points for different domains. The models using the logistic weighted data perform better than those using the unweighted data. The different definitions of positive ALS do not make a large difference in terms of detection of onset. We believe that the onset of activity limitation for adults is most likely to occur between ages 50 and 60.

Age	Education		
	Low	Median	High
40	0.043	**0.415**	0.023
41	0.053	0.322	0.066
42	0.053	0.084	0.019
43	0.101	0.057	0.08
44	0.119	0.011	**0.249**
45	0.146	0.007	0.12
46	0.187	0.021	0.092
47	**0.193**	0.011	0.079
48	0.064	0.019	0.031
49	0.014	0.017	0.037
50	0.009	0.019	0.026
51	0.004	0.013	0.024
52	0.011	0.002	0.022
53	0	0.002	0.03
54	0.001	0	0.018
55	0.001	0	0.009
56	0	0	0.011
57	0	0	0.013
58	0	0	0.025
59	0	0	0.016
60	0	0	0.004
61	0	0	0.004
62	0	0	0.001
63	0	0	0
64	0.001	0	0
65	0	0	0
66	0	0	0
67	0	0	0
68	0	0	0
69	0	0	0
70	0	0	0.001

Table 5.2: Distributions for the change point k for the 3 domains using the revised Bayesian hierachical model and unweighted data, where only "Limited in major activity" is considered as positive ALS.

Age	Education		
	Low	Median	High
40	0.007	**0.448**	0.007
41	0.011	0.442	0.125
42	0.028	0.054	0.003
43	0.057	0.041	0.152
44	0.186	0.001	**0.232**
45	**0.314**	0.001	0.227
46	0.226	0.003	0.1
47	0.147	0.002	0.049
48	0.023	0.001	0.017
49	0	0.003	0.036
50	0	0.004	0.04
51	0	0	0.009
52	0	0	0.001
53	0	0	0.001
54	0.001	0	0
55	0	0	0.001
56	0	0	0
57	0	0	0
58	0	0	0
59	0	0	0
60	0	0	0
61	0	0	0
62	0	0	0
63	0	0	0
64	0	0	0
65	0	0	0
66	0	0	0
67	0	0	0
68	0	0	0
69	0	0	0
70	0	0	0

Table 5.3: Distributions for the change point k for the 3 domains using the revised Bayesian hierachical model and unweighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Age	Education		
	Low	Median	High
40	0.01	0.069	0
41	0.02	0.061	0
42	0.02	0.069	0
43	0.029	**0.138**	0
44	0.042	0.084	0
45	0.055	0.104	0
46	0.058	0.072	0
47	0.069	0.048	0
48	0.066	0.061	0
49	0.066	0.056	0.002
50	0.09	0.036	0.002
51	**0.109**	0.043	0.002
52	0.088	0.024	0.012
53	0.064	0.016	0.013
54	0.048	0.029	0.029
55	0.05	0.015	0.027
56	0.036	0.026	0.038
57	0.029	0.023	0.06
58	0.024	0.01	0.073
59	0.009	0.002	0.106
60	0.009	0.009	**0.14**
61	0.005	0.003	0.128
62	0.003	0.001	0.087
63	0	0.001	0.068
64	0	0	0.047
65	0	0	0.049
66	0	0	0.037
67	0	0	0.027
68	0.001	0	0.022
69	0	0	0.023
70	0	0	0.008

Table 5.4: Distributions for the change point k for the 3 domains using the revised Bayesian hierachical model and logistic weighted data, where only "Limited in major activity" is considered as positive ALS.

Age	Education		
	Low	Median	High
40	0.001	0.018	0
41	0	0.023	0
42	0.002	0.024	0.003
43	0.004	0.034	0
44	0.005	0.045	0.005
45	0.013	0.046	0.011
46	0.013	0.056	0.013
47	0.026	0.059	0.03
48	0.036	**0.082**	0.032
49	0.04	**0.082**	0.042
50	0.058	0.065	0.059
51	0.055	0.059	0.082
52	0.059	0.051	0.09
53	0.082	0.065	0.096
54	0.091	0.047	**0.119**
55	**0.094**	0.04	0.09
56	0.087	0.036	0.094
57	0.082	0.045	0.052
58	0.089	0.036	0.046
59	0.082	0.023	0.046
60	0.05	0.021	0.023
61	0.013	0.017	0.023
62	0.01	0.009	0.014
63	0.006	0.004	0.016
64	0.001	0.004	0.01
65	0	0.005	0.003
66	0	0.003	0
67	0	0	0
68	0	0.001	0.001
69	0.001	0	0
70	0	0	0

Table 5.5: Distributions for the change point k for the 3 domains using the revised Bayesian hierachical model and logistic weighted data, where both "Limited in major activity" and "Limited in kind/amount of major activity" are considered as positive ALS.

Bibliography

[1] P.F. Adams and M.A. Marano. Current estimates from the national health interview survey, 1994. national center for health statistics. hyattsville, md. *Vital Health Statistics*, Series 10(No. 193), December 1995.

[2] T. Burchardt. Being and becoming: Social exclusion and the onset of disability. Technical Report CASEreport21, London: Centre for Analysis of Social Exclusion, London School of Economics, 2003.

[3] B.P. Carlin, A.E. Gelfand, and M. Smith, A.F. Hierarchical bayesian analysis of change point problems. *Applied Statistics*, 41:389–405, 1992.

[4] D.R. Cox. *Analysis of Binary Data*. Chapman and Hall, 1970.

[5] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via em algorithm. *Journal of the Royal Statistical Society*, 39(B):1–38, 1977.

[6] J. Girón, J. Ginebra, and A. Riba. Bayesian analysis of a multinomial sequence and homogeneity of literary style. *The American Statistician*, 59(1):19–30, Feburary 2005.

[7] P.J. Green. Reversible jump markov chain monte carlo computation and bayesian model determination. *Biometrika*, 82:711–732, 1995.

[8] K. Kalton and I. Flores-Cervantes. Weighting methods. *Journal of Official Statistics*, 19(2):81–97, 2003.

[9] A.F.M. Smith. A bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 63:407–416, 1975.

[10] Z. Zimmer and J. S. House. Education, income and functional limitation trasitions among american adults: Contrasting onset and progression. *International Journal of Epidemiology*, 32(6):1089–1097, 2003.